INTRODUCTION TO ECONOMETRIC THEORY

CHUNG-MING KUAN

Institute of Economics
Academia Sinica

This version: September 15, 2000

 Address for correspondence: Institute of Economics, Academia Sinica, Taipei 115, Taiwan; e-mail: ckuan@ieas.econ.sinica.edu.tw.

Contents

Chapter 1

Linear and Matrix Algebra

This chapter summarizes some important results of linear and matrix algebra that are instrumental in analyzing the statistical methods in subsequent chapters. The coverage of these mathematical topics is rather brief but self-contained. Readers may also consult other linear and matrix algebra textbooks for more detailed discussions; see e.g., Anton (1981), Basilevsky (1983), Graybill (1969), and Noble and Daniel (1977).

In this chapter we first introduce basic matrix notations (Section 1.1) and matrix operations (Section 1.2). We then study the determinant and trace functions (Section 1.3), matrix inverse (Section 1.4), and matrix rank (Section 1.5). After introducing eigenvalue and diagonalization (Section 1.6), we discuss the properties of symmetric matrix (Section 1.7) and orthogonal projection in a vector space (Section 1.8).

1.1 Basic Notations

A *matrix* is an array of numbers. In what follows, a matrix is denoted by an upper-case alphabet in boldface (e.g., $\boldsymbol{A}$), and its $(i,j)^{\text{th}}$ element (the element at the i^{th} row and j^{th} column) is denoted by the corresponding lower-case alphabet with subscripts ij (e.g., a_{ij}). Specifically, a $m \times n$ matrix $\boldsymbol{A}$ contains m rows and n columns and can be expressed as

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

An $n \times 1$ ($1 \times n$) matrix is an n-dimensional column (row) *vector*. Every vector will be denoted by a lower-case alphabet in boldface (e.g., $\boldsymbol{z}$), and its i^{th} element is denoted

by the corresponding lower-case alphabet with subscript i (e.g., z_i). An 1×1 matrix is just a *scalar*. For a matrix $\boldsymbol{A}$, its i^{th} column is denoted as $\boldsymbol{a}_i$.

A matrix is *square* if its number of rows equals the number of columns. A matrix is said to be *diagonal* if its off-diagonal elements (i.e., a_{ij}, $i \neq j$) are all zeros and at least one of its diagonal elements is non-zero, i.e., $a_{ii} \neq 0$ for some $i = 1, \ldots, n$. A diagonal matrix whose diagonal elements are all ones is an *identity* matrix, denoted as $\boldsymbol{I}$; we also write the $n \times n$ identity matrix as $\boldsymbol{I}_n$. A matrix $\boldsymbol{A}$ is said to be *lower* (*upper*) *triangular* if $a_{ij} = 0$ for $i < (>)\, j$. We let $\boldsymbol{0}$ denote the matrix whose elements are all zeros.

For a vector-valued function $\boldsymbol{f} : \mathbb{R}^m \to \mathbb{R}^n$, $\nabla_{\boldsymbol{\theta}} \boldsymbol{f}(\boldsymbol{\theta})$ is the $m \times n$ matrix of the first-order derivatives of $\boldsymbol{f}$ with respect to the elements of $\boldsymbol{\theta}$:

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{f}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial f_1(\boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial f_2(\boldsymbol{\theta})}{\partial \theta_1} & \cdots & \frac{\partial f_n(\boldsymbol{\theta})}{\partial \theta_1} \\ \frac{\partial f_1(\boldsymbol{\theta})}{\partial \theta_2} & \frac{\partial f_2(\boldsymbol{\theta})}{\partial \theta_2} & \cdots & \frac{\partial f_n(\boldsymbol{\theta})}{\partial \theta_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1(\boldsymbol{\theta})}{\partial \theta_m} & \frac{\partial f_2(\boldsymbol{\theta})}{\partial \theta_m} & \cdots & \frac{\partial f_n(\boldsymbol{\theta})}{\partial \theta_m} \end{bmatrix}.$$

When $n = 1$, $\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$ is the (column) *gradient* vector of $f(\boldsymbol{\theta})$. The $m \times m$ *Hessian* matrix of the second-order derivatives of the real-valued function $f(\boldsymbol{\theta})$ is

$$\nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} (\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})) = \begin{bmatrix} \frac{\partial^2 f(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_1} & \frac{\partial^2 f(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 f(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_m} \\ \frac{\partial^2 f(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 f(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_2} & \cdots & \frac{\partial^2 f(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\boldsymbol{\theta})}{\partial \theta_m \partial \theta_1} & \frac{\partial^2 f(\boldsymbol{\theta})}{\partial \theta_m \partial \theta_2} & \cdots & \frac{\partial^2 f(\boldsymbol{\theta})}{\partial \theta_m \partial \theta_m} \end{bmatrix}.$$

1.2 Matrix Operations

Two matrices are said to be of the same size if they have the same number of rows and same number of columns. Matrix equality is defined for two matrices of the same size. Given two $m \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} = \boldsymbol{B}$ if $a_{ij} = b_{ij}$ for every i, j. The *transpose* of an $m \times n$ matrix $\boldsymbol{A}$, denoted as $\boldsymbol{A}'$, is the $n \times m$ matrix whose $(i, j)^{\text{th}}$ element is the $(j, i)^{\text{th}}$ element of $\boldsymbol{A}$. The transpose of a column vector is a row vector; the transpose of a scalar is just the scalar itself. A matrix $\boldsymbol{A}$ is said to be *symmetric* if $\boldsymbol{A} = \boldsymbol{A}'$, i.e., $a_{ij} = a_{ji}$ for all i, j. Clearly, a diagonal matrix is symmetric, but a triangular matrix is not.

Matrix addition is also defined for two matrices of the same size. Given two $m \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, their sum, $\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B}$, is the $m \times n$ matrix with the $(i, j)^{\text{th}}$ element

$c_{ij} = a_{ij} + b_{ij}$. Note that matrix addition, if defined, is commutative:

$$\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{B} + \boldsymbol{A},$$

and associative:

$$\boldsymbol{A} + (\boldsymbol{B} + \boldsymbol{C}) = (\boldsymbol{A} + \boldsymbol{B}) + \boldsymbol{C}.$$

Also, $\boldsymbol{A} + \boldsymbol{0} = \boldsymbol{A}$.

The scalar multiplication of the scalar c and matrix $\boldsymbol{A}$ is the matrix $c\boldsymbol{A}$ whose $(i,j)^{\text{th}}$ element is ca_{ij}. Clearly, $c\boldsymbol{A} = \boldsymbol{A}c$, and $-\boldsymbol{A} = -1 \times \boldsymbol{A}$. Thus, $\boldsymbol{A} + (-\boldsymbol{A}) = \boldsymbol{A} - \boldsymbol{A} = \boldsymbol{0}$. Given two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, the matrix multiplication $\boldsymbol{AB}$ is defined only when the number of columns of $\boldsymbol{A}$ is the same as the number of rows of $\boldsymbol{B}$. Specifically, when $\boldsymbol{A}$ is $m \times n$ and $\boldsymbol{B}$ is $n \times p$, their product, $\boldsymbol{C} = \boldsymbol{AB}$, is the $m \times p$ matrix whose $(i,j)^{\text{th}}$ element is

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}.$$

Matrix multiplication is not commutative, i.e., $\boldsymbol{AB} \neq \boldsymbol{BA}$; in fact, when $\boldsymbol{AB}$ is defined, $\boldsymbol{BA}$ need not be defined. On the other hand, matrix multiplication is associative:

$$\boldsymbol{A}(\boldsymbol{BC}) = (\boldsymbol{AB})\boldsymbol{C},$$

and distributive with respect to matrix addition:

$$\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}.$$

It is easy to verify that $(\boldsymbol{AB})' = \boldsymbol{B}'\boldsymbol{A}'$. For an $m \times n$ matrix $\boldsymbol{A}$, $\boldsymbol{I}_m \boldsymbol{A} = \boldsymbol{A}\boldsymbol{I}_n = \boldsymbol{A}$.

The *inner product* of two d-dimensional vectors $\boldsymbol{y}$ and $\boldsymbol{z}$ is the scalar

$$\boldsymbol{y}'\boldsymbol{z} = \sum_{i=1}^{d} y_i z_i.$$

If $\boldsymbol{y}$ is m-dimensional and $\boldsymbol{z}$ is n-dimensional, their *outer product* is the matrix $\boldsymbol{y}\boldsymbol{z}'$ whose $(i,j)^{\text{th}}$ element is $y_i z_j$. In particular,

$$\boldsymbol{z}'\boldsymbol{z} = \sum_{i=1}^{d} z_i^2,$$

which is non-negative and induces the standard *Euclidean norm* of $\boldsymbol{z}$ as $\|\boldsymbol{z}\| = (\boldsymbol{z}'\boldsymbol{z})^{1/2}$. The vector with Euclidean norm zero must be a zero vector; the vector with Euclidean norm one is referred to as a *unit vector*. For example,

$$(1\ 0\ 0), \qquad \Big(0\ \frac{1}{2}\ \frac{\sqrt{3}}{2}\Big), \qquad \Big(\frac{1}{\sqrt{2}}\ \frac{1}{\sqrt{3}}\ \frac{1}{\sqrt{6}}\Big),$$

are all unit vectors. A vector whose i^{th} element is one and the remaining elements are all zero is called the i^{th} Cartesian unit vector.

Let θ denote the angle between $\boldsymbol{y}$ and $\boldsymbol{z}$. By the law of cosine,

$$\|\boldsymbol{y}-\boldsymbol{z}\|^2 = \|\boldsymbol{y}\|^2 + \|\boldsymbol{z}\|^2 - 2\|\boldsymbol{y}\|\,\|\boldsymbol{z}\|\cos\theta,$$

where the left-hand side is $\|\boldsymbol{y}\|^2 + \|\boldsymbol{z}\|^2 - 2\boldsymbol{y}'\boldsymbol{z}$. Thus, the inner product of $\boldsymbol{y}$ and $\boldsymbol{z}$ can be expressed as

$$\boldsymbol{y}'\boldsymbol{z} = \|\boldsymbol{y}\|\|\boldsymbol{z}\|\cos\theta.$$

When $\theta = \pi/2$, $\cos\theta = 0$ so that $\boldsymbol{y}'\boldsymbol{z} = 0$. In this case, we say that $\boldsymbol{y}$ and $\boldsymbol{z}$ are *orthogonal* to each other. A square matrix $\boldsymbol{A}$ is said to be *orthogonal* if $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}' = \boldsymbol{I}$. Hence, each column (row) vector of an orthogonal matrix is a unit vector and orthogonal to all remaining column (row) vectors. When $\boldsymbol{y} = c\boldsymbol{z}$ for some $c \neq 0$, $\theta = 0$ or π, and $\boldsymbol{y}$ and $\boldsymbol{z}$ are said to be *linearly dependent*.

As $-1 \leq \cos\theta \leq 1$, we immediately obtain the so-called *Cauchy-Schwarz inequality*.

Lemma 1.1 (Cauchy-Schwarz) *For two d-dimensional vectors $\boldsymbol{y}$ and $\boldsymbol{z}$,*

$$|\boldsymbol{y}'\boldsymbol{z}| \leq \|\boldsymbol{y}\|\|\boldsymbol{z}\|,$$

where the equality holds when $\boldsymbol{y}$ and $\boldsymbol{z}$ are linearly dependent.

It follows from the Cauchy-Schwarz inequality that

$$\begin{aligned}\|\boldsymbol{y}+\boldsymbol{z}\|^2 &= \|\boldsymbol{y}\|^2 + \|\boldsymbol{z}\|^2 + 2\boldsymbol{y}'\boldsymbol{z} \\ &\leq \|\boldsymbol{y}\|^2 + \|\boldsymbol{z}\|^2 + 2\|\boldsymbol{y}\|\|\boldsymbol{z}\| \\ &= (\|\boldsymbol{y}\| + \|\boldsymbol{z}\|)^2.\end{aligned}$$

This leads to the following *triangle inequality*.

Lemma 1.2 *For two d-dimensional vectors $\boldsymbol{y}$ and $\boldsymbol{z}$,*

$$\|\boldsymbol{y}+\boldsymbol{z}\| \leq \|\boldsymbol{y}\| + \|\boldsymbol{z}\|,$$

where the equality holds when $\boldsymbol{y} = c\boldsymbol{z}$ for some $c > 0$.

When $\boldsymbol{y}$ and $\boldsymbol{z}$ are orthogonal,

$$\|\boldsymbol{y}+\boldsymbol{z}\|^2 = \|\boldsymbol{y}\|^2 + \|\boldsymbol{z}\|^2,$$

which is the celebrated *Pythagoras theorem*.

A special type of matrix multiplication, known as the *Kronecker product*, is defined for matrices without size restrictions. Specifically, the Kronecker product of two matrices $\boldsymbol{A}$ $(m \times n)$ and $\boldsymbol{B}$ $(p \times q)$ is the $mp \times nq$ matrix:

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2n}\boldsymbol{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}\boldsymbol{B} & a_{m2}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{bmatrix}.$$

The Kronecker product is not commutative:

$$\boldsymbol{A} \otimes \boldsymbol{B} \neq \boldsymbol{B} \otimes \boldsymbol{A},$$

but it is associative:

$$(\boldsymbol{A} \otimes \boldsymbol{B}) \otimes \boldsymbol{C} = \boldsymbol{A} \otimes (\boldsymbol{B} \otimes \boldsymbol{C}).$$

It also obeys the distributive law:

$$\boldsymbol{A} \otimes (\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{A} \otimes \boldsymbol{B} + \boldsymbol{A} \otimes \boldsymbol{C}.$$

It can be verified that

$$(\boldsymbol{A} \otimes \boldsymbol{B})' = \boldsymbol{A}' \otimes \boldsymbol{B}'.$$

Consider now differentiation involving vectors and matrices. Let $\boldsymbol{a}$ and $\boldsymbol{\theta}$ be tow d-dimensional vectors. We have

$$\nabla_{\boldsymbol{\theta}} (\boldsymbol{a}'\boldsymbol{\theta}) = \boldsymbol{a}.$$

For a symmetric matrix $\boldsymbol{A}$,

$$\nabla_{\boldsymbol{\theta}} (\boldsymbol{\theta}'\boldsymbol{A}\boldsymbol{\theta}) = 2\boldsymbol{A}\boldsymbol{\theta}, \qquad \nabla^2_{\boldsymbol{\theta}} (\boldsymbol{\theta}'\boldsymbol{A}\boldsymbol{\theta}) = 2\boldsymbol{A}.$$

1.3 Matrix Determinant and Trace

Given a square matrix $\boldsymbol{A}$, let $\boldsymbol{A}_{ij}$ denote the sub-matrix obtained from $\boldsymbol{A}$ by deleting its i^{th} row and j^{th} column. The *determinant* of $\boldsymbol{A}$ is

$$\det(\boldsymbol{A}) = \sum_{i=1}^{m} (-1)^{i+j} a_{ij} \det(\boldsymbol{A}_{ij}),$$

for any $j = 1, \ldots, n$, where $(-1)^{i+j} \det(\boldsymbol{A}_{ij})$ is called the *cofactor* of a_{ij}. This definition is based on the cofactor expansion along the j^{th} column. Equivalently, the determinant can also be defined using the cofactor expansion along the i^{th} row:

$$\det(\boldsymbol{A}) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(\boldsymbol{A}_{ij}),$$

for any $i = 1, \ldots, m$. The determinant of a scalar is the scalar itself; the determinant of a 2×2 matrix $\boldsymbol{A}$ is simply $a_{11}a_{22} - a_{12}a_{21}$. A square matrix with non-zero determinant is said to be *nonsingular*; otherwise, it is *singular*.

Clearly, $\det(\boldsymbol{A}) = \det(\boldsymbol{A}')$. From the definition of determinant, it is straightforward to see that for a scalar c and an $n \times n$ matrix $\boldsymbol{A}$,

$$\det(c\boldsymbol{A}) = c^n \det(\boldsymbol{A}),$$

and that for a square matrix with a column (or row) of zeros, its determinant must be zero. Also, the determinant of a diagonal or triangular matrix is simply the product of all the diagonal elements. It can also be shown that the determinant of the product of two square matrices of the same size is the product of their determinants:

$$\det(\boldsymbol{AB}) = \det(\boldsymbol{A}) \det(\boldsymbol{B}) = \det(\boldsymbol{BA}).$$

Also, for an $m \times m$ matrix $\boldsymbol{A}$ and a $p \times p$ matrix $\boldsymbol{B}$,

$$\det(\boldsymbol{A} \otimes \boldsymbol{B}) = \det(\boldsymbol{A})^m \det(\boldsymbol{B})^p.$$

If $\boldsymbol{A}$ is an orthogonal matrix, we know $\boldsymbol{AA}' = \boldsymbol{I}$ so that

$$\det(\boldsymbol{I}) = \det(\boldsymbol{AA}') = [\det(\boldsymbol{A})]^2.$$

As the determinant of the identity matrix is one, the determinant of an orthogonal matrix must be either 1 or -1.

The *trace* of a square matrix is the sum of its diagonal elements; i.e., $\text{trace}(\boldsymbol{A}) = \sum_i a_{ii}$. For example, $\text{trace}(\boldsymbol{I}_n) = n$. Clearly, $\text{trace}(\boldsymbol{A}) = \text{trace}(\boldsymbol{A}')$. The trace function has the linear property:

$$\text{trace}(c\boldsymbol{A} + d\boldsymbol{B}) = c\,\text{trace}(\boldsymbol{A}) + d\,\text{trace}(\boldsymbol{B}),$$

where c and d are scalars. It can also be shown that

$$\text{trace}(\boldsymbol{AB}) = \text{trace}(\boldsymbol{BA}),$$

provided that both $\boldsymbol{AB}$ and $\boldsymbol{BA}$ are defined. For two square matrices $\boldsymbol{A}$ and$\boldsymbol{B}$,

$$\text{trace}(\boldsymbol{A} \otimes \boldsymbol{B}) = \text{trace}(\boldsymbol{A})\,\text{trace}(\boldsymbol{B}).$$

1.4 Matrix Inverse

A nonsingular matrix $\boldsymbol{A}$ possesses a unique *inverse* $\boldsymbol{A}^{-1}$ in the sense that $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$. A singular matrix cannot be inverted, however. Thus, saying that a matrix is invertible is equivalent to saying that it is nonsingular.

Given an invertible matrix $\boldsymbol{A}$, its inverse can be calculated as

$$\boldsymbol{A}^{-1} = \frac{1}{\det(\boldsymbol{A})}\boldsymbol{F}',$$

where $\boldsymbol{F}$ is the matrix of cofactors, i.e., the $(i,j)^{\text{th}}$ element of $\boldsymbol{F}$ is the cofactor $(-1)^{i+j}\det(\boldsymbol{A}_{ij})$. The matrix $\boldsymbol{F}'$ is known as the *adjoint* of $\boldsymbol{A}$. For example, when $\boldsymbol{A}$ is 2×2,

$$\boldsymbol{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}.$$

Matrix inversion and transposition can be interchanged, i.e., $(\boldsymbol{A}')^{-1} = (\boldsymbol{A}^{-1})'$. For two nonsingular matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same size, we have $\boldsymbol{A}\boldsymbol{B}\boldsymbol{B}^{-1}\boldsymbol{A}^{-1} = \boldsymbol{I}$, so that

$$(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}.$$

Some special matrices can be easily inverted. For example, for a diagonal matrix $\boldsymbol{A}$, $\boldsymbol{A}^{-1}$ is also diagonal with the diagonal elements a_{ii}^{-1}; for an orthogonal matrix $\boldsymbol{A}$, $\boldsymbol{A}^{-1} = \boldsymbol{A}'$.

A formula for computing the inverse of a partitioned matrix is

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{F}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} & -\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{F}^{-1} \\ -\boldsymbol{F}^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} & \boldsymbol{F}^{-1} \end{bmatrix},$$

where $\boldsymbol{F} = \boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}$, or equivalently,

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{G}^{-1} & -\boldsymbol{G}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{G}^{-1} & \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{G}^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \end{bmatrix},$$

where $\boldsymbol{G} = \boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}$, provided that the matrix inverses in the expressions above are well defined. In particular, if this matrix is *block diagonal* so that the off-diagonal blocks are zero matrices, we have

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{A}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}^{-1} \end{bmatrix},$$

provided that $\boldsymbol{A}$ and $\boldsymbol{D}$ are invertible.

1.5 Matrix Rank

The vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are said to be *linearly independent* if the only solution to

$$c_1\boldsymbol{z}_1 + c_2\boldsymbol{z}_2 + \cdots + c_n\boldsymbol{z}_n = 0$$

is the trivial solution: $c_1 = \cdots = c_n = 0$; otherwise, they are *linearly dependent*. When two (three) vectors are linearly dependent, they are on the same line (plane).

The *column* (*row*) *rank* of a matrix $\boldsymbol{A}$ is the maximum number of linearly independent column (row) vectors of $\boldsymbol{A}$. When the column (row) rank equals the number of column (row) vectors, this matrix is said to be of full column (row) rank. The space *spanned* by the vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ is the collection of all linear combinations of these vectors, denoted as $\text{span}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$. The space spanned by the column vectors of $\boldsymbol{A}$ is therefore $\text{span}(\boldsymbol{A})$, which is also known as the *column space* of $\boldsymbol{A}$. A vector $\boldsymbol{z}$ is in $\text{span}(\boldsymbol{A})$ if it can be expressed as $\boldsymbol{Ac}$ for some vector $\boldsymbol{c} \neq \boldsymbol{0}$. Similarly, the space spanned by the row vectors of $\boldsymbol{A}$ is $\text{span}(\boldsymbol{A}')$ and known as the *row space* of $\boldsymbol{A}$. The column (row) rank of $\boldsymbol{A}$ is the *dimension* of the column (row) space of $\boldsymbol{A}$.

Given an $n \times k$ matrix $\boldsymbol{A}$ with $k \leq n$, suppose that $\boldsymbol{A}$ has row rank $r \leq n$ and column rank $c \leq k$. Without loss of generality, assume that the first r row vectors are linear independent. Hence, each row vector $\boldsymbol{a}_i$ can be expressed as

$$\boldsymbol{a}_i = q_{i1}\boldsymbol{a}_1 + q_{i2}\boldsymbol{a}_2 + \cdots + q_{ir}\boldsymbol{a}_r, \qquad i = 1, \ldots, n,$$

with the j^{th} element

$$a_{ij} = q_{i1}a_{1j} + q_{i2}a_{2j} + \cdots + q_{ir}a_{rj}, \qquad i = 1, \ldots, n, \;\; j = 1, \ldots, k.$$

Fixing j, we immediately see that every column vector of $\boldsymbol{A}$ can be written as a linear combination of the vectors $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_r$. As such, the column rank of $\boldsymbol{A}$ must be less than or equal to r. Similarly, the column rank of $\boldsymbol{A}'$, which is also the row rank of $\boldsymbol{A}$, must be less than or equal to c. This proves the following result.

Lemma 1.3 *The column rank and row rank of a matrix are equal.*

By Lemma 1.3, we can then define the *rank* of $\boldsymbol{A}$ as the maximum number of linearly independent column (or row) vectors of $\boldsymbol{A}$, denoted as $\text{rank}(\boldsymbol{A})$. Clearly, $\text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{A}')$. An $n \times n$ matrix $\boldsymbol{A}$ is said to be of *full rank* if $\text{rank}(\boldsymbol{A}) = n$.

For an $n \times k$ matrix $\boldsymbol{A}$, its *left inverse* is a $k \times n$ matrix $\boldsymbol{A}_L^{-1}$ such that $\boldsymbol{A}_L^{-1}\boldsymbol{A} = \boldsymbol{I}_k$. Similarly, a *right inverse* of $\boldsymbol{A}$ is a $k \times n$ matrix $\boldsymbol{A}_R^{-1}$ such that $\boldsymbol{A}\boldsymbol{A}_R^{-1} = \boldsymbol{I}_n$. The left

and right inverses are not unique, however. It can be shown that a matrix possesses a left (right) inverse if and only if it has full column (row) rank. Thus, for a square matrix with full rank, it has both inverses, which are just the unique matrix inverse. Thus, a nonsingular (invertible) matrix must be of full rank and vice versa.

It can be shown that for two $n \times k$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,

$$\text{rank}(\boldsymbol{A}+\boldsymbol{B}) \leq \text{rank}(\boldsymbol{A}) + \text{rank}(\boldsymbol{B}).$$

If $\boldsymbol{A}$ is $n \times k$ and $\boldsymbol{B}$ is $k \times m$,

$$\text{rank}(\boldsymbol{A}) + \text{rank}(\boldsymbol{B}) - k \leq \text{rank}(\boldsymbol{AB}) \leq \min[\text{rank}(\boldsymbol{A}), \text{rank}(\boldsymbol{B})].$$

For the Kronecker product, we have

$$\text{rank}(\boldsymbol{A} \otimes \boldsymbol{B}) = \text{rank}(\boldsymbol{A})\,\text{rank}(\boldsymbol{B}).$$

If $\boldsymbol{A}$ is a nonsingular matrix, we have from the inequality above that

$$\text{rank}(\boldsymbol{AB}) \leq \text{rank}(\boldsymbol{B}) = \text{rank}(\boldsymbol{A}^{-1}\boldsymbol{AB}) \leq \text{rank}(\boldsymbol{AB});$$

i.e., $\text{rank}(\boldsymbol{AB}) = \text{rank}(\boldsymbol{B})$. This also implies that for a nonsingular matrix $\boldsymbol{C}$,

$$\text{rank}(\boldsymbol{BC}) = \text{rank}(\boldsymbol{C}'\boldsymbol{B}') = \text{rank}(\boldsymbol{B}') = \text{rank}(\boldsymbol{B}).$$

Thus, the rank of a matrix is preserved under nonsingular transformations.

Lemma 1.4 *Let $\boldsymbol{A}$ $(n \times n)$ and $\boldsymbol{C}$ $(k \times k)$ be nonsingular matrices. Then for any $n \times k$ matrix $\boldsymbol{B}$,*

$$\text{rank}(\boldsymbol{B}) = \text{rank}(\boldsymbol{AB}) = \text{rank}(\boldsymbol{BC}).$$

1.6 Eigenvalue and Eigenvector

Given a square matrix $\boldsymbol{A}$, if $\boldsymbol{Ac} = \lambda\boldsymbol{c}$ for some scalar λ and non-zero vector $\boldsymbol{c}$, then $\boldsymbol{c}$ is an *eigenvector* of $\boldsymbol{A}$ corresponding to the *eigenvalue* λ. The system $(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{c} = \boldsymbol{0}$ has a non-trivial solution if and only if

$$\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0.$$

This is known as the *characteristic equation* of $\boldsymbol{A}$, from which we can solve for the eigenvalues of $\boldsymbol{A}$. Hence, eigenvalues (eigenvectors) are also referred to as *characteristic roots* (*characteristic vectors*). Note that the eigenvalues and eigenvectors of a real-valued matrix need not be real-valued.

When $\boldsymbol{A}$ is $n \times n$, the characteristic equation is an n^{th}-order polynomial in λ and has at most n distinct solutions. These solutions (eigenvalues) are usually complex-valued. If some eigenvalues take the same value, there may exist several eigenvectors corresponding to the same eigenvalue. Given an eigenvalue λ, let $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k$ be associated eigenvectors. Then,

$$\boldsymbol{A}(a_1\boldsymbol{c}_1 + a_2\boldsymbol{c}_2 + \ldots + a_k\boldsymbol{c}_k) = \lambda(a_1\boldsymbol{c}_1 + a_2\boldsymbol{c}_2 + \ldots + a_k\boldsymbol{c}_k),$$

so that any linear combination of these eigenvectors is again an eigenvector corresponding to λ. That is, these eigenvectors are closed under scalar multiplication and vector addition and form the *eigenspace* corresponding to λ. As such, for a common eigenvalue, we are mainly concerned with those eigenvectors that are linearly independent.

If $\boldsymbol{A}$ $(n \times n)$ possesses n distinct eigenvalues, each eigenvalue must correspond to one eigenvector, unique up to scalar multiplications. It is therefore typical to normalize eigenvectors such that they have Euclidean norm one. It can also be shown that if the eigenvalues of a matrix are all distinct, their associated eigenvectors must be linearly independent. Let $\boldsymbol{C}$ denote the matrix of these eigenvectors and $\boldsymbol{\Lambda}$ denote the diagonal matrix with diagonal elements being the eigenvalues of $\boldsymbol{A}$. We can write $\boldsymbol{AC} = \boldsymbol{C\Lambda}$. As $\boldsymbol{C}$ is nonsingular, we have

$$\boldsymbol{C}^{-1}\boldsymbol{AC} = \boldsymbol{\Lambda}, \qquad \text{or} \qquad \boldsymbol{A} = \boldsymbol{C\Lambda C}^{-1}.$$

In this case, $\boldsymbol{A}$ is said to be *similar* to $\boldsymbol{\Lambda}$.

When $\boldsymbol{A}$ has n distinct eigenvalues $\lambda_1, \ldots, \lambda_n$, it is readily seen that

$$\det(\boldsymbol{A}) = \det(\boldsymbol{C\Lambda C}^{-1}) = \det(\boldsymbol{\Lambda})\det(\boldsymbol{C})\det(\boldsymbol{C}^{-1}) = \det(\boldsymbol{\Lambda}),$$

and

$$\operatorname{trace}(\boldsymbol{A}) = \operatorname{trace}(\boldsymbol{C\Lambda C}^{-1}) = \operatorname{trace}(\boldsymbol{C}^{-1}\boldsymbol{C\Lambda}) = \operatorname{trace}(\boldsymbol{\Lambda}).$$

This yields the following result.

Lemma 1.5 *Let $\boldsymbol{A}$ be an $n \times n$ matrix with distinct eigenvalues $\lambda_1, \ldots, \lambda_n$. Then*

$$\det(\boldsymbol{A}) = \det(\boldsymbol{\Lambda}) = \prod_{i=1}^{n} \lambda_i,$$

$$\operatorname{trace}(\boldsymbol{A}) = \operatorname{trace}(\boldsymbol{\Lambda}) = \sum_{i=1}^{n} \lambda_i.$$

When $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}^{-1}$, we have $\boldsymbol{A}^{-1} = \boldsymbol{C}\boldsymbol{\Lambda}^{-1}\boldsymbol{C}^{-1}$. This shows that the eigenvectors of $\boldsymbol{A}^{-1}$ are the same as those of $\boldsymbol{A}$, and the corresponding eigenvalues are simply the reciprocals of the eigenvalues of $\boldsymbol{A}$. Similarly,

$$\boldsymbol{A}^2 = (\boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}^{-1})(\boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}^{-1}) = \boldsymbol{C}\boldsymbol{\Lambda}^2\boldsymbol{C}^{-1},$$

so that the eigenvectors of $\boldsymbol{A}^2$ are the same as those of $\boldsymbol{A}$, and the corresponding eigenvalues are the squares of the eigenvalues of $\boldsymbol{A}$. This result generalizes immediately to $\boldsymbol{A}^k$.

1.7 Symmetric Matrix

More can be said about symmetric matrices. Let $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ be two eigenvectors of $\boldsymbol{A}$ corresponding to the distinct eigenvalues λ_1 and λ_2, respectively. If $\boldsymbol{A}$ is symmetric, then

$$\boldsymbol{c}_2'\boldsymbol{A}\boldsymbol{c}_1 = \lambda_1\boldsymbol{c}_2'\boldsymbol{c}_1 = \lambda_2\boldsymbol{c}_2'\boldsymbol{c}_1.$$

As $\lambda_1 \neq \lambda_2$, it must be true that $\boldsymbol{c}_2'\boldsymbol{c}_1 = 0$, so that they are orthogonal. Given linearly independent eigenvectors that correspond to a common eigenvalue, they can also be orthogonalized. Thus, a symmetric matrix is *orthogonally diagonalizable*, in the sense that

$$\boldsymbol{C}'\boldsymbol{A}\boldsymbol{C} = \boldsymbol{\Lambda}, \qquad \text{or} \qquad \boldsymbol{A} = \boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}',$$

where $\boldsymbol{\Lambda}$ is again the diagonal matrix of the eigenvalues of $\boldsymbol{A}$, and $\boldsymbol{C}$ is the orthogonal matrix of associated eigenvectors.

As nonsingular transformations preserve rank (Lemma 1.4), so do orthogonal transformations. We thus have the result below.

Lemma 1.6 *For a symmetric matrix $\boldsymbol{A}$, $\operatorname{rank}(\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{\Lambda})$, the number of non-zero eigenvalues of $\boldsymbol{A}$.*

Moreover, when $\boldsymbol{A}$ is diagonalizable, the assertions of Lemma 1.5 remain valid, whether or not the eigenvalues of $\boldsymbol{A}$ are distinct.

Lemma 1.7 *Let $\boldsymbol{A}$ be an $n \times n$ symmetric matrix. Then,*

$$\det(\boldsymbol{A}) = \det(\boldsymbol{\Lambda}) = \prod_{i=1}^{n} \lambda_i,$$

$$\operatorname{trace}(\boldsymbol{A}) = \operatorname{trace}(\boldsymbol{\Lambda}) = \sum_{i=1}^{n} \lambda_i.$$

By Lemma 1.7, a symmetric matrix is nonsingular if its eigenvalues are all non-zero.

A symmetric matrix $\boldsymbol{A}$ is said to be *positive definite* if $\boldsymbol{b}'\boldsymbol{A}\boldsymbol{b} > 0$ for all vectors $\boldsymbol{b} \neq \boldsymbol{0}$; $\boldsymbol{A}$ is said to be *positive semi-definite* if $\boldsymbol{b}'\boldsymbol{A}\boldsymbol{b} \geq 0$ for all $\boldsymbol{b} \neq \boldsymbol{0}$. A positive definite matrix thus must be nonsingular, but a positive semi-definite matrix may be singular. Suppose that $\boldsymbol{A}$ is a symmetric matrix orthogonally diagonalized as $\boldsymbol{C}'\boldsymbol{A}\boldsymbol{C} = \boldsymbol{\Lambda}$. If $\boldsymbol{A}$ is also positive semi-definite, then for any $\boldsymbol{b} \neq \boldsymbol{0}$,

$$\boldsymbol{b}'\boldsymbol{\Lambda}\boldsymbol{b} = \boldsymbol{b}'(\boldsymbol{C}'\boldsymbol{A}\boldsymbol{C})\boldsymbol{b} = \tilde{\boldsymbol{b}}'\boldsymbol{A}\tilde{\boldsymbol{b}} \geq 0,$$

where $\tilde{\boldsymbol{b}} = \boldsymbol{C}\boldsymbol{b}$. This shows that $\boldsymbol{\Lambda}$ is also positive semi-definite, and all the diagonal elements of $\boldsymbol{\Lambda}$ must be non-negative. It can be seen that the converse also holds.

Lemma 1.8 *A symmetric matrix is positive definite (positive semi-definite) if, and only if, its eigenvalues are all positive (non-negative).*

For a symmetric and positive definite matrix $\boldsymbol{A}$, $\boldsymbol{A}^{-1/2}$ is such that $\boldsymbol{A}^{-1/2\prime}\boldsymbol{A}^{-1/2} = \boldsymbol{A}^{-1}$. In particular, by orthogonal diagonalization,

$$\boldsymbol{A}^{-1} = \boldsymbol{C}\boldsymbol{\Lambda}^{-1}\boldsymbol{C}' = (\boldsymbol{C}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{C}')(\boldsymbol{C}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{C}'),$$

so that we may choose $\boldsymbol{A}^{-1/2} = \boldsymbol{C}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{C}'$. The inverse of $\boldsymbol{A}^{-1/2}$ is $\boldsymbol{A}^{1/2} = \boldsymbol{C}\boldsymbol{\Lambda}^{1/2}\boldsymbol{C}'$. It follows that $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2\prime} = \boldsymbol{A}$, and $\boldsymbol{A}^{-1/2}\boldsymbol{A}\boldsymbol{A}^{-1/2\prime} = \boldsymbol{I}$. Note that $\boldsymbol{\Lambda}^{-1/2}\boldsymbol{C}'$ is also a legitimate choice of $\boldsymbol{A}^{-1/2}$, yet it is not symmetric.

Finally, we know that for two positive real numbers a and b, $a \geq b$ implies $b^{-1} \geq a^{-1}$. This result can be generalized to compare two positive definite matrices, as stated below without proof.

Lemma 1.9 *Given two symmetric and positive definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, if $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite, then so is $\boldsymbol{B}^{-1} - \boldsymbol{A}^{-1}$.*

1.8 Orthogonal Projection

A matrix $\boldsymbol{A}$ is said to be *idempotent* if $\boldsymbol{A}^2 = \boldsymbol{A}$. Given a vector $\boldsymbol{y}$ in the Euclidean space V, a *projection* of $\boldsymbol{y}$ onto a subspace S of V is a linear transformation of $\boldsymbol{y}$ to S. The resulting projected vector can be written as $\boldsymbol{P}\boldsymbol{y}$, where $\boldsymbol{P}$ is the associated transformation matrix. Given the projection $\boldsymbol{P}\boldsymbol{y}$ in S, further projection to S should have no effect on $\boldsymbol{P}\boldsymbol{y}$, i.e.,

$$\boldsymbol{P}(\boldsymbol{P}\boldsymbol{y}) = \boldsymbol{P}^2\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y}.$$

Thus, a matrix $\boldsymbol{P}$ is said to be a *projection matrix* if it is *idempotent*.

A projection of $\boldsymbol{y}$ onto S is orthogonal if the projection $\boldsymbol{P}\boldsymbol{y}$ is orthogonal to the difference between $\boldsymbol{y}$ and $\boldsymbol{P}\boldsymbol{y}$. That is,

$$(\boldsymbol{y} - \boldsymbol{P}\boldsymbol{y})'\boldsymbol{P}\boldsymbol{y} = \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P})'\boldsymbol{P}\boldsymbol{y} = \boldsymbol{0}.$$

As $\boldsymbol{y}$ is arbitrary, the equality above holds if, and only if, $(\boldsymbol{I} - \boldsymbol{P})'\boldsymbol{P} = \boldsymbol{0}$. Consequently, $\boldsymbol{P} = \boldsymbol{P}'\boldsymbol{P}$ and $\boldsymbol{P}' = \boldsymbol{P}'\boldsymbol{P}$. This shows that $\boldsymbol{P}$ must be symmetric. Thus, a matrix is an *orthogonal projection matrix* if, and only if, it is symmetric and idempotent. It can be easily verified that the orthogonal projection $\boldsymbol{P}\boldsymbol{y}$ must be unique.

When $\boldsymbol{P}$ is an orthogonal projection matrix, it is easily seen that $\boldsymbol{I} - \boldsymbol{P}$ is idempotent because

$$(\boldsymbol{I} - \boldsymbol{P})^2 = \boldsymbol{I} - 2\boldsymbol{P} + \boldsymbol{P}^2 = \boldsymbol{I} - \boldsymbol{P}.$$

As $\boldsymbol{I} - \boldsymbol{P}$ is also symmetric, it is an orthogonal projection matrix. Since $(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{P} = \boldsymbol{0}$, the projections $\boldsymbol{P}\boldsymbol{y}$ and $(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}$ must be orthogonal. This shows that any vector $\boldsymbol{y}$ can be uniquely decomposed into two orthogonal components:

$$\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y} + (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}.$$

Define the *orthogonal complement* of a subspace $S \subseteq V$ as

$$S^\perp = \{\boldsymbol{v} \in V : \boldsymbol{v}'\boldsymbol{s} = 0, \text{ for all } \boldsymbol{s} \in S\}.$$

If $\boldsymbol{P}$ is the orthogonal projection matrix that projects vectors onto $S \subseteq V$, we have $\boldsymbol{P}\boldsymbol{s} = \boldsymbol{s}$ for any $\boldsymbol{s} \in S$. It follows that $(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}$ is orthogonal to $\boldsymbol{s}$ and that $(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}$ is the orthogonal projection of $\boldsymbol{y}$ onto $S^\perp$.

Intuitively, the orthogonal projection $\boldsymbol{Py}$ can be interpreted as the "best approximation" of $\boldsymbol{y}$ in S, in the sense that $\boldsymbol{Py}$ is the closest to $\boldsymbol{y}$ in terms of the Euclidean norm. To see this, we observe that for any $\boldsymbol{s} \in S$,

$$\begin{aligned}\|\boldsymbol{y}-\boldsymbol{s}\|^2 &= \|\boldsymbol{y}-\boldsymbol{Py}+\boldsymbol{Py}-\boldsymbol{s}\|^2 \\ &= \|\boldsymbol{y}-\boldsymbol{Py}\|^2+\|\boldsymbol{Py}-\boldsymbol{s}\|^2+2(\boldsymbol{y}-\boldsymbol{Py})'(\boldsymbol{Py}-\boldsymbol{s}) \\ &= \|\boldsymbol{y}-\boldsymbol{Py}\|^2+\|\boldsymbol{Py}-\boldsymbol{s}\|^2.\end{aligned}$$

This establishes the following result.

Lemma 1.10 *Let $\boldsymbol{y}$ be a vector in V and $\boldsymbol{Py}$ its orthogonal projection onto $S \subseteq V$. Then,*

$$\|\boldsymbol{y}-\boldsymbol{Py}\| \leq \|\boldsymbol{y}-\boldsymbol{s}\|,$$

for all $\boldsymbol{s} \in S$.

Let $\boldsymbol{A}$ be a symmetric and idempotent matrix and $\boldsymbol{C}$ be the orthogonal matrix that diagonalizes $\boldsymbol{A}$ to $\boldsymbol{\Lambda}$. Then,

$$\boldsymbol{\Lambda} = \boldsymbol{C}'\boldsymbol{AC} = \boldsymbol{C}'\boldsymbol{A}(\boldsymbol{CC}')\boldsymbol{AC} = \boldsymbol{\Lambda}^2.$$

This is possible only when the eigenvalues of $\boldsymbol{A}$ are zero and one. The result below now follows from Lemmas 1.8.

Lemma 1.11 *A symmetric and idempotent matrix is positive semi-definite with the eigenvalues 0 and 1.*

Moreover, $\operatorname{trace}(\boldsymbol{\Lambda})$ is the number of non-zero eigenvalues of $\boldsymbol{A}$ and hence $\operatorname{rank}(\boldsymbol{\Lambda})$. When $\boldsymbol{A}$ is symmetric, $\operatorname{rank}(\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{\Lambda})$ by Lemma 1.6, and $\operatorname{trace}(\boldsymbol{A}) = \operatorname{trace}(\boldsymbol{\Lambda})$ by Lemma 1.7. Combining these results we have:

Lemma 1.12 *For a symmetric and idempotent matrix $\boldsymbol{A}$, $\operatorname{rank}(\boldsymbol{A}) = \operatorname{trace}(\boldsymbol{A})$, the number of non-zero eigenvalues of $\boldsymbol{A}$.*

Given an $n \times k$ matrix $\boldsymbol{A}$, it is easy to see that $\boldsymbol{A}'\boldsymbol{A}$ and $\boldsymbol{AA}'$ are symmetric and positive semi-definite. Let $\boldsymbol{x}$ denote a vector orthogonal to the rows of $\boldsymbol{A}'\boldsymbol{A}$; i.e., $\boldsymbol{A}'\boldsymbol{Ax} = \boldsymbol{0}$. Hence $\boldsymbol{x}'\boldsymbol{A}'\boldsymbol{Ax} = 0$, so that $\boldsymbol{Ax}$ must be a zero vector. That is, $\boldsymbol{x}$ is also orthogonal to the rows of $\boldsymbol{A}$. Conversely, $\boldsymbol{Ax} = \boldsymbol{0}$ implies $\boldsymbol{A}'\boldsymbol{Ax} = \boldsymbol{0}$. This shows

that the orthogonal complement of the row space of $\boldsymbol{A}$ is the same as the orthogonal complement of the row space of $\boldsymbol{A}'\boldsymbol{A}$. Hence, these two row spaces are also the same. Similarly, the column space of $\boldsymbol{A}$ is the same as the column space of $\boldsymbol{A}\boldsymbol{A}'$. It follows from Lemma 1.3 that

$$\text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{A}'\boldsymbol{A}) = \text{rank}(\boldsymbol{A}\boldsymbol{A}').$$

In particular, if $\boldsymbol{A}$ $(n \times k)$ is of full column rank $k < n$, then $\boldsymbol{A}'\boldsymbol{A}$ is $k \times k$ and hence of full rank k (nonsingular), but $\boldsymbol{A}\boldsymbol{A}'$ is $n \times n$ and hence singular. The result below is now immediate.

Lemma 1.13 *If $\boldsymbol{A}$ is an $n \times k$ matrix with full column rank $k < n$, then, $\boldsymbol{A}'\boldsymbol{A}$ is symmetric and positive definite.*

Given an $n \times k$ matrix $\boldsymbol{A}$ with full column rank $k < n$, $\boldsymbol{P} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$ is clearly symmetric and idempotent and hence an orthogonal projection matrix. As

$$\text{trace}(\boldsymbol{P}) = \text{trace}(\boldsymbol{A}'\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}) = \text{trace}(\boldsymbol{I}_k) = k,$$

we have from Lemmas 1.11 and 1.12 that $\boldsymbol{P}$ has exactly k eigenvalues equal to 1 and that $\text{rank}(\boldsymbol{P}) = k$. Similarly, $\text{rank}(\boldsymbol{I} - \boldsymbol{P}) = n - k$. Moreover, any vector $\boldsymbol{y} \in \text{span}(\boldsymbol{A})$ can be written as $\boldsymbol{A}\boldsymbol{b}$ for some non-zero vector $\boldsymbol{b}$, and

$$\boldsymbol{P}\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'(\boldsymbol{A}\boldsymbol{b}) = \boldsymbol{A}\boldsymbol{b} = \boldsymbol{y}.$$

This suggests that $\boldsymbol{P}$ must project vectors onto $\text{span}(\boldsymbol{A})$. On the other hand, when $\boldsymbol{y} \in \text{span}(\boldsymbol{A})^{\perp}$, $\boldsymbol{y}$ is orthogonal to the column vectors of $\boldsymbol{A}$ so that $\boldsymbol{A}'\boldsymbol{y} = \boldsymbol{0}$. It follows that $\boldsymbol{P}\boldsymbol{y} = \boldsymbol{0}$ and $(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y} = \boldsymbol{y}$. Thus, $\boldsymbol{I} - \boldsymbol{P}$ must project vectors onto $\text{span}(\boldsymbol{A})^{\perp}$. These results are summarized below.

Lemma 1.14 *Let $\boldsymbol{A}$ be an $n \times k$ matrix with full column rank k. Then, $\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$ orthogonally projects vectors onto* $\text{span}(\boldsymbol{A})$ *and has rank k; $\boldsymbol{I}_n - \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$ orthogonally projects vectors onto* $\text{span}(\boldsymbol{A})^{\perp}$ *and has rank $n - k$.*

References

Anton, Howard (1981). *Elementary Linear Algebra*, third edition, New York: Wiley.

Basilevsky, Alexander (1983). *Applied Matrix Algebra in the Statistical Sciences*, New York: North-Holland.

Graybill, Franklin A. (1969). *Introduction to Matrices with Applications in Statistics*, second edition, Belmont, CA: Wadsworth.

Noble, Ben and James W. Daniel. *Applied Linear Algebra*, second edition, Englewood Cliffs, NJ: Prentice-Hall.

Chapter 2

Statistical Concepts

In this chapter we summarize some basic probability and statistics results to be used in subsequent chapters. We focus on finite-sample results of multivariate random vectors and statistics; asymptotic properties require more profound mathematical tools and will not be discussed until Chapter 5. The topics covered in this chapter can be found in most of statistics textbooks; in particular, Amemiya (1994) is a useful reference.

2.1 Distribution Functions

Given a random experiment, let Ω denote the collection of all possible outcomes of this experiment and $\mathbb{P}$ denote the probability measure assigned to a certain collection of events (subsets of Ω). If A is an event, $\mathbb{P}(A)$ is such that $0 \leq \mathbb{P}(A) \leq 1$ and measures the likelihood of A. The larger is $\mathbb{P}(A)$, the more likely is the event A to occur. A d-dimensional random vector ($\mathbb{R}^d$-valued random variable) is a function of the outcomes $\omega \in \Omega$ and takes values in $\mathbb{R}^d$. Formal definitions of probability space and random variables are given in Section 5.1.

The (joint) *distribution function* of the $\mathbb{R}^d$-valued random variable $\boldsymbol{z}$ is the non-decreasing, right-continuous function $F_{\boldsymbol{z}}$ such that for $\boldsymbol{\zeta} = (\zeta_1 \ \cdots \ \zeta_d)' \in \mathbb{R}^d$,

$$F_{\boldsymbol{z}}(\boldsymbol{\zeta}) = \mathbb{P}\{\omega \in \Omega \colon z_1(\omega) \leq \zeta_1, \ldots, z_d(\omega) \leq \zeta_d\},$$

with

$$\lim_{\zeta_1 \to -\infty, \ldots, \zeta_d \to -\infty} F_{\boldsymbol{z}}(\boldsymbol{\zeta}) = 0, \qquad \lim_{\zeta_1 \to \infty, \ldots, \zeta_d \to \infty} F_{\boldsymbol{z}}(\boldsymbol{\zeta}) = 1.$$

Note that the distribution function of $\boldsymbol{z}$ is a standard point function defined on $\mathbb{R}^d$ and provides a convenient way to characterize the randomness of $\boldsymbol{z}$. The (joint) *density*

function of $F_{\boldsymbol{z}}$, if exists, is the non-negative function $f_{\boldsymbol{z}}$ such that

$$F_{\boldsymbol{z}}(\boldsymbol{\zeta}) = \int_{-\infty}^{\zeta_d} \cdots \int_{-\infty}^{\zeta_1} f_{\boldsymbol{z}}(s_1, \ldots, s_d)\, \mathrm{d}s_1 \cdots \mathrm{d}s_d,$$

where the right-hand side is a Riemann integral. Clearly, the density function $f_{\boldsymbol{z}}$ must be integrated to one on $\mathbb{R}^d$.

The *marginal distribution function* of the i^{th} component of $\boldsymbol{z}$ is

$$F_{z_i}(\zeta_i) = \mathbb{P}\{\omega \in \Omega \colon z_i(\omega) \leq \zeta_i\} = F_{\boldsymbol{z}}(\infty, \ldots, \infty, \zeta_i, \infty, \ldots, \infty).$$

Thus, the marginal distribution function of z_i is the joint distribution function without restrictions on the other elements z_j, $j \neq i$. The *marginal density function* of z_i is the non-negative function f_{z_i} such that

$$F_{z_i}(\zeta_i) = \int_{-\infty}^{\zeta_i} f_{z_i}(s)\, \mathrm{d}s.$$

It is readily seen that the marginal density function f_{z_i} can also be obtained from the associated joint density function by integrating out the other elements:

$$f_{z_i}(s_i) = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f_{\boldsymbol{z}}(s_1, \ldots, s_d)\, \mathrm{d}s_1 \cdots \mathrm{d}s_{i-1}\, \mathrm{d}s_{i+1} \cdots \mathrm{d}s_d.$$

If there are two random vectors $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, they are said to be *independent* if, and only if, their joint distribution function is the product of all marginal distribution functions:

$$F_{\boldsymbol{z}_1, \boldsymbol{z}_2}(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2) = F_{\boldsymbol{z}_1}(\boldsymbol{\zeta}_1)\, F_{\boldsymbol{z}_2}(\boldsymbol{\zeta}_2);$$

otherwise, they are *dependent*. If random vectors possess density functions, they are independent if, and only if, their joint density function is also the product of marginal density functions. It is intuitively clear that functions of independent random variables remain independent, as stated in the result below.

Lemma 2.1 *If $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are independent random vectors, then their transformations, $h_1(\boldsymbol{z}_1)$ and $h_2(\boldsymbol{z}_2)$, are also independent random variables.*

2.2 Moments

Given the d-dimensional random vector $\boldsymbol{z}$ with the distribution function $F_{\boldsymbol{z}}$, the expectation of the i^{th} element z_i is defined as

$$\mathbb{E}(z_i) = \int_{\mathbb{R}^d} \cdots \int \zeta_i\, \mathrm{d}F_{\boldsymbol{z}}(\zeta_1, \ldots, \zeta_d),$$

where the right-hand side is a Stieltjes integral; for more details about different integrals we refer to Rudin (1976). As this integral equals

$$\int_{\mathbb{R}} \zeta_i \, \mathrm{d}F_{\boldsymbol{z}}(\infty, \ldots, \infty, \zeta_i, \infty, \ldots, \infty) = \int_{\mathbb{R}} \zeta_i \, \mathrm{d}F_{z_i}(\zeta_i),$$

the expectation of z_i can be taken with respect to either the joint distribution function $F_{\boldsymbol{z}}$ or the marginal distribution function F_{z_i}.

We say that the random variable z_i has a finite expected value (or the expectation $\mathbb{E}(z_i)$ exists) if $\mathbb{E}\,|z_i| < \infty$. A random variable need not have a finite expected value; if it does, this random variable is said to be *integrable*. More generally, the expectation of a random vector is defined elementwise. Thus, for a random vector $\boldsymbol{z}$, $\mathbb{E}(\boldsymbol{z})$ exists if all $\mathbb{E}(z_i)$, $i = 1, \ldots, d$, exist, and $\boldsymbol{z}$ is integrable if all z_i, $i = 1, \ldots, d$, are integrable.

It is easily seen that the expectation operator does not have any effect on a constant; that is, $\mathbb{E}(b) = b$ for any constant b. For integrable random variables z_i and z_j, the expectation operator is *monotonic* in the sense that

$$\mathbb{E}(z_i) \leq \mathbb{E}(z_j),$$

for any $z_i \leq z_j$ with probability one. Moreover, the expectation operator possesses the *linearity* property:

$$\mathbb{E}(az_i + bz_j) = a\,\mathbb{E}(z_i) + b\,\mathbb{E}(z_j),$$

where a and b are two real numbers. This property immediately generalizes to integrable random vectors.

Lemma 2.2 *Let $\boldsymbol{A}$ $(n \times d)$ and $\boldsymbol{B}$ $(n \times c)$ be two non-stochastic matrices. Then for any integrable random vectors $\boldsymbol{z}$ $(d \times 1)$ and $\boldsymbol{y}$ $(c \times 1)$,*

$$\mathbb{E}(\boldsymbol{A}\boldsymbol{z} + \boldsymbol{B}\boldsymbol{y}) = \boldsymbol{A}\,\mathbb{E}(\boldsymbol{z}) + \boldsymbol{B}\,\mathbb{E}(\boldsymbol{y}).$$

If $\boldsymbol{b}$ is an n-dimensional nonstochastic vector, then $\mathbb{E}(\boldsymbol{A}\boldsymbol{z} + \boldsymbol{b}) = \boldsymbol{A}\,\mathbb{E}(\boldsymbol{z}) + \boldsymbol{b}$.

More generally, let $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{z})$ be a well-defined, vector-valued function of $\boldsymbol{z}$. The expectation of $\boldsymbol{y}$ is

$$\mathbb{E}(\boldsymbol{y}) = \mathbb{E}[\boldsymbol{g}(\boldsymbol{z})] = \int_{\mathbb{R}^d} \boldsymbol{g}(\boldsymbol{\zeta}) \; \mathrm{d}F_{\boldsymbol{z}}(\boldsymbol{\zeta}).$$

When $\boldsymbol{g}(\boldsymbol{z}) = z_i^k$, $\mathbb{E}[\boldsymbol{g}(\boldsymbol{z})] = \mathbb{E}(z_i^k)$ is known as the k^{th} *moment* of z_i, where k need not be an integer. In particular, $\mathbb{E}(z_i)$ is the first moment of z_i. When a random variable

has finite k^{th} moment, its moments of order less than k are also finite. Thus, if the k^{th} moment does not exist, then the moments of order greater than k also fail to exist. See Section 2.3 for some examples of random variables that possess only low order moments. A random vector is said to have finite k^{th} moment if its elements all have finite k^{th} moment. A random variable with finite second moment is said to be *square integrable*; a random vector is square integrable if its elements are all square integrable.

The k^{th} *central moment* of z_i is $\mathbb{E}[z_i - \mathbb{E}(z_i)]^k$. In particular, the second central moment of the square integrable random variable z_i is

$$\mathbb{E}[z_i - \mathbb{E}(z_i)]^2 = \mathbb{E}(z_i^2) - [\mathbb{E}(z_i)]^2,$$

which is a measure of dispersion of the values of z_i. The second central moment is also known as *variance*, denoted as $\text{var}(\cdot)$. The square root of variance is *standard deviation*. It can be verified that, given the square integrable random variable z_i and real numbers a and b,

$$\text{var}(az_i + b) = \text{var}(az_i) = a^2 \text{ var}(z_i).$$

This shows that variance is location invariant but not scale invariant.

When $g(\boldsymbol{z}) = z_i z_j$, $\mathbb{E}[g(\boldsymbol{z})] = \mathbb{E}(z_i z_j)$ is the *cross moment* of z_i and z_j. The *cross central moment* of z_i and z_j is

$$\mathbb{E}[(z_i - \mathbb{E}(z_i))(z_j - \mathbb{E}(z_j))] = \mathbb{E}(z_i z_j) - \mathbb{E}(z_i)\,\mathbb{E}(z_j),$$

which is a measure of the co-variation between these two random variables. The cross central moment of two random variables is known as their *covariance*, denoted as $\text{cov}(\cdot,\cdot)$. Clearly, $\text{cov}(z_i, z_j) = \text{cov}(z_j, z_i)$ and $\text{cov}(z_i, z_i) = \text{var}(z_i)$. It can be seen that for real numbers a, b, c, d,

$$\text{cov}(az_i + b, cz_j + d) = \text{cov}(az_i, cz_j) = ac\,\text{cov}(z_i, z_j).$$

Thus, covariance is also location invariant but depends on the scale (measurement units) of random variables.

Observe that for any real numbers a and b,

$$\text{var}(az_i + bz_j) = a^2 \text{ var}(z_i) + b^2 \text{ var}(z_j) + 2ab \text{ cov}(z_i, z_j),$$

so that

$$\text{var}(z_i - az_j) = \text{var}(z_i) + a^2\,\text{var}(z_j) - 2a\,\text{cov}(z_i, z_j),$$

which must be non-negative. Setting $a = \operatorname{cov}(z_i, z_j)/\operatorname{var}(z_j)$, we have

$$\operatorname{var}(z_i) - \operatorname{cov}(z_i, z_j)^2/\operatorname{var}(z_j) \geq 0.$$

In particular, when $z_i = az_j + b$ for some real numbers a and b, we have $\operatorname{var}(z_i) = a^2 \operatorname{var}(z_j)$ and $\operatorname{cov}(z_i, z_j) = a \operatorname{var}(z_j)$, so that

$$\operatorname{var}(z_i) - \operatorname{cov}(z_i, z_j)^2/\operatorname{var}(z_j) = 0.$$

This yields the Cauchy-Schwarz inequality for square integrable random variables.

Lemma 2.3 (Cauchy-Schwarz) *Let z_i, z_j be two square integrable random variables. Then,*

$$\operatorname{cov}(z_i, z_j)^2 \leq \operatorname{var}(z_i) \operatorname{var}(z_j),$$

where the equality holds when $z_i = az_j + b$ for some real numbers a and b.

cf. the Cauchy-Schwarz inequality (Lemma 1.1) in Section 1.2. This also suggests that when two random variables are square integrable, their covariance must be finite.

The *correlation coefficient* of z_i and z_j is defined as

$$\operatorname{corr}(z_i, z_j) = \frac{\operatorname{cov}(z_i, z_j)}{\sqrt{\operatorname{var}(z_i) \operatorname{var}(z_j)}}.$$

By Lemma 2.3 we have

$$-1 \leq \operatorname{corr}(z_i, z_j) \leq 1.$$

If $\operatorname{corr}(z_i, z_j) = 0$, z_i and z_j are said to be *uncorrelated*. If $\operatorname{corr}(z_i, z_j) > 0$, z_i and z_j are said to be positively correlated; if $\operatorname{corr}(z_i, z_j) < 0$, z_i and z_j are negatively correlated. When $z_i = az_j + b$, $\operatorname{corr}(z_i, z_j) = 1$ if $a > 0$ and -1 if $a < 0$. In both cases, z_i and z_j are perfectly correlated. For two random variables z_i and z_j and real numbers a, b, c, d,

$$\operatorname{corr}(az_i + b, cz_j + d) = \operatorname{corr}(az_i, cz_j) = \frac{ac}{|a|\,|c|} \operatorname{corr}(z_i, z_j).$$

Thus, the correlation coefficient is not only location invariant but also scale invariant, apart from the sign change.

For a d-dimensional, square integrable random vector $\boldsymbol{z}$, its variance-covariance matrix is

$$\begin{aligned}\mathrm{var}(\boldsymbol{z}) &= \mathbb{E}[(\boldsymbol{z}-\mathbb{E}(\boldsymbol{z}))(\boldsymbol{z}-\mathbb{E}(\boldsymbol{z}))'] \\ &= \begin{bmatrix} \mathrm{var}(z_1) & \mathrm{cov}(z_1,z_2) & \cdots & \mathrm{cov}(z_1,z_d) \\ \mathrm{cov}(z_2,z_1) & \mathrm{var}(z_2) & \cdots & \mathrm{cov}(z_2,z_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(z_d,z_1) & \mathrm{cov}(z_d,z_2) & \cdots & \mathrm{var}(z_d) \end{bmatrix}.\end{aligned}$$

As $\mathrm{cov}(z_i,z_j)=\mathrm{cov}(z_j,z_i)$, $\mathrm{var}(\boldsymbol{z})$ must be symmetric. Moreover, $\mathrm{var}(\boldsymbol{z})$ is positive semi-definite because it is the expectation of a matrix that is positive semi-definite with probability one.

For two random vectors $\boldsymbol{y}$ $(c\times 1)$ and $\boldsymbol{z}$ $(d\times 1)$, the $d\times c$ covariance matrix of $\boldsymbol{z}$ and $\boldsymbol{y}$ is

$$\mathrm{cov}(\boldsymbol{z},\boldsymbol{y}) = \mathbb{E}[(\boldsymbol{z}-\mathbb{E}\,\boldsymbol{z})(\boldsymbol{y}-\mathbb{E}\,\boldsymbol{y})'] = \mathbb{E}(\boldsymbol{z}\boldsymbol{y}') - \mathbb{E}(\boldsymbol{z})\,\mathbb{E}(\boldsymbol{y}').$$

Two random vectors are uncorrelated if their covariance matrix is a zero matrix. If $\boldsymbol{y}$ and $\boldsymbol{z}$ are independent, their joint distribution function is the product of individual distribution functions. It follows that the cross moment of $\boldsymbol{y}$ and $\boldsymbol{z}$ is the product of their individual first moment: that

$$\mathbb{E}(\boldsymbol{z}\boldsymbol{y}') = \mathbb{E}(\boldsymbol{z})\,\mathbb{E}(\boldsymbol{y}').$$

This shows that independence implies $\mathrm{cov}(\boldsymbol{z},\boldsymbol{y})=\boldsymbol{0}$. Uncorrelated random vectors are not necessarily independent, however.

Based on the properties of variance and covariance for random variables, we have the following result for random vectors.

Lemma 2.4 *Let $\boldsymbol{A}$ $(n\times d)$, $\boldsymbol{B}$ $(n\times c)$, and $\boldsymbol{C}$ $(m\times c)$ be non-stochastic matrices and $\boldsymbol{b}$ an n-dimensional non-stochastic vector. Then for any square integrable random vectors $\boldsymbol{z}$ $(d\times 1)$ and $\boldsymbol{y}$ $(c\times 1)$,*

$$\begin{aligned}\mathrm{var}(\boldsymbol{A}\boldsymbol{z}+\boldsymbol{B}\boldsymbol{y}) &= \boldsymbol{A}\,\mathrm{var}(\boldsymbol{z})\boldsymbol{A}' + \boldsymbol{B}\,\mathrm{var}(\boldsymbol{y})\boldsymbol{B}' + 2\boldsymbol{A}\,\mathrm{cov}(\boldsymbol{z},\boldsymbol{y})\boldsymbol{B}', \\ \mathrm{var}(\boldsymbol{A}\boldsymbol{z}+\boldsymbol{b}) &= \mathrm{var}(\boldsymbol{A}\boldsymbol{z}) = \boldsymbol{A}\,\mathrm{var}(\boldsymbol{z})\,\boldsymbol{A}'.\end{aligned}$$

Given two square integrable random vectors $\boldsymbol{z}$ and $\boldsymbol{y}$, suppose that $\mathrm{var}(\boldsymbol{y})$ is positive definite. As the variance-covariance matrix of $(\boldsymbol{z}'\ \boldsymbol{y}')'$ must be a positive semi-definite matrix,

$$\begin{aligned}[I\ -\mathrm{cov}(\boldsymbol{z},\boldsymbol{y})\,\mathrm{var}(\boldsymbol{y})^{-1}] &\begin{bmatrix} \mathrm{var}(\boldsymbol{z}) & \mathrm{cov}(\boldsymbol{z},\boldsymbol{y}) \\ \mathrm{cov}(\boldsymbol{y},\boldsymbol{z}) & \mathrm{var}(\boldsymbol{y}) \end{bmatrix} \begin{bmatrix} I \\ -\mathrm{var}(\boldsymbol{y})^{-1}\,\mathrm{cov}(\boldsymbol{y},\boldsymbol{z}) \end{bmatrix} \\ &= \mathrm{var}(\boldsymbol{z}) - \mathrm{cov}(\boldsymbol{z},\boldsymbol{y})\,\mathrm{var}(\boldsymbol{y})^{-1}\,\mathrm{cov}(\boldsymbol{y},\boldsymbol{z})\end{aligned}$$

is also a positive semi-definite matrix. This establishes the multivariate version of the Cauchy-Schwarz inequality for square integrable random vectors.

Lemma 2.5 (Cauchy-Schwarz) *Let $\boldsymbol{y},\boldsymbol{z}$ be two square integrable random vectors. Then,*

$$\mathrm{var}(\boldsymbol{z}) - \mathrm{cov}(\boldsymbol{z},\boldsymbol{y})\,\mathrm{var}(\boldsymbol{y})^{-1}\,\mathrm{cov}(\boldsymbol{y},\boldsymbol{z})$$

is a positive semi-definite matrix.

A random vector is said to be *degenerate* (have a singular distribution) if its variance-covariance matrix is singular. Let $\boldsymbol{\Sigma}$ be the variance-covariance matrix of the d-dimensional random vector $\boldsymbol{z}$. If $\boldsymbol{\Sigma}$ is singular, then there exists a non-zero vector $\boldsymbol{c}$ such that $\boldsymbol{\Sigma c} = \boldsymbol{0}$. For this particular $\boldsymbol{c}$, we have

$$\boldsymbol{c}'\boldsymbol{\Sigma c} = \mathbb{E}[\boldsymbol{c}'(\boldsymbol{z} - \mathbb{E}(\boldsymbol{z}))]^2 = 0.$$

It follows that $\boldsymbol{c}'[\boldsymbol{z} - \mathbb{E}(\boldsymbol{z})] = 0$ with probability one; i.e, the elements of $\boldsymbol{z}$ are linearly dependent with probability one. This implies that all the probability mass of $\boldsymbol{z}$ is concentrated in a subspace of dimension less than d.

2.3 Special Distributions

In this section we discuss the multivariate normal (Gaussian) distribution and other univariate distributions such as the chi-square, Student's t, and Fisher's F distributions.

A random vector $\boldsymbol{z}$ is said to have a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, denoted as $\boldsymbol{z} \sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$, if it has the density function

$$\frac{1}{(2\pi)^{d/2}\,\det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right).$$

For $d = 1$, this is just the density of the univariate normal random variable. Note that the multivariate normal density function is completely characterized by its mean vector and variance-covariance matrix. A normal random variable has moments of all orders; in particular, its even-order central moments are

$$\mathbb{E}(z-\mu)^k = (k-1)\cdots 3\cdot 1\ \mathrm{var}(z)^{k/2}, \qquad k \geq 2 \text{ and } k \text{ is even},$$

and its odd-order central moments are all zeros. A normal random variable with mean zero and variance one is usually called the standard normal random variable.

When $\boldsymbol{\Sigma}$ is a diagonal matrix with diagonal elements σ_{ii}, $i = 1, \ldots, d$, the elements of $\boldsymbol{z}$ are uncorrelated. Note that for normal random variables, uncorrelatedness implies independence. In this case, the density function is simply the product of marginal density functions for $z_1, \ldots, z_d$:

$$\frac{1}{(2\pi)^{d/2}(\prod_{i=1}^d \sigma_{ii})^{1/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^d \frac{(z_i-\mu_i)^2}{\sigma_{ii}}\right).$$

When $\sigma_{ii} = \sigma_o^2$, a constant, this joint density simplifies to

$$\frac{1}{(2\pi\sigma_o^2)^{d/2}} \exp\left(-\frac{1}{2\sigma_o^2}\sum_{i=1}^d (z_i-\mu_i)^2\right).$$

Although uncorrelated normal random variables are also independent, we stress again that this need not be true for other random variables.

The result below shows that proper linear transformations of normal random vectors remain normally distributed.

Lemma 2.6 *Let $\boldsymbol{z}$ be a d-dimensional random vector distributed as $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Also let $\boldsymbol{A}$ be an $n \times d$ non-stochastic matrix with full row rank $n < d$ and $\boldsymbol{b}$ be a d-dimensional non-stochastic vector. Then,*

$$\boldsymbol{A}\boldsymbol{z} + \boldsymbol{b} \sim N(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}').$$

Lemma 2.6 implies that, when $\boldsymbol{z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, any sub-vector (element) of $\boldsymbol{z}$ also has a multivariate (univariate) normal distribution; the converse need not be true, however. It is also easily seen that

$$\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{z} - \boldsymbol{\mu}) \sim N(\boldsymbol{0}, \boldsymbol{I}_d),$$

where $\boldsymbol{\Sigma}^{-1/2}$ is such that $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{I}$, as defined in Section 1.7. Proper standardization of a normal random vector thus yields a normal random vector with independent elements. If $\boldsymbol{A}$ is not of full row rank, $\text{var}(\boldsymbol{A}\boldsymbol{z}) = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'$ does not have full rank, so that $\boldsymbol{A}\boldsymbol{z}$ is degenerate.

Let $\boldsymbol{z} \sim N(\boldsymbol{\mu}, \boldsymbol{I}_d)$. The sum of squares of the elements of $\boldsymbol{z}$ is the *non-central chi-square* random variable with d degrees of freedom and the *non-centrality parameter* $\nu = \boldsymbol{\mu}'\boldsymbol{\mu}$, denoted as

$$\boldsymbol{z}'\boldsymbol{z} \sim \chi^2(d;\, \nu).$$

The density function of $\chi^2(d;\, \nu)$ is

$$f(x) = \exp\Big(-\frac{\nu + x}{2}\Big) x^{d/2-1} \frac{1}{2^{d/2}} \sum_{i=0}^{\infty} \frac{x^i \nu^i}{i!\, 2^{2i}\, \Gamma(i + d/2)}, \qquad x > 0,$$

where Γ is the gamma function with

$$\Gamma(n) = \int_0^\infty e^{-x} x^{n-1}\, \mathrm{d}x.$$

It can be shown that a $\chi^2(d;\, \nu)$ random variable has mean $(d+\nu)$ and variance $2d+4\nu$. When $\boldsymbol{\mu} = \boldsymbol{0}$, the non-centrality parameter $\nu = 0$, and $\chi^2(d;\, 0)$ is known as the central chi-square random variable, denoted as $\chi^2(d)$. The density of $\chi^2(d)$ is

$$f(x) = \exp\Big(-\frac{x}{2}\Big) x^{d/2-1} \frac{1}{2^{d/2}\, \Gamma(d/2)}, \qquad x > 0,$$

with mean d and variance $2d$. The result below follows directly from Lemma 2.6.

Lemma 2.7 *Let $\boldsymbol{z}$ be a d-dimensional random vector distributed as $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then,*

$$\boldsymbol{z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z} \sim \chi^2(d;\, \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu});$$

in particular, if $\boldsymbol{\mu} = \boldsymbol{0}$, $\boldsymbol{z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z} \sim \chi^2(d)$.

Let w and x be two independent random variables such that $w \sim N(\mu, 1)$ and $x \sim \chi^2(n)$. Then

$$\frac{w}{\sqrt{x/n}} \sim t(n;\, \mu),$$

the non-central t distribution with n degrees of freedom and the non-centrality parameter μ. The density function of $t(n;\, \mu)$ is

$$f(x) = \frac{n^{n/2} \exp(-\mu^2/2)}{\Gamma(n/2)\Gamma(1/2)(n + x^2)^{(n+1)/2}} \sum_{i=0}^{\infty} \Gamma\Big(\frac{n+i+1}{2}\Big) \frac{\mu^i}{i!} \left(\frac{2x^2}{n + x^2}\right)^{i/2} (\text{sign } x)^i.$$

When $\mu = 0$, $t(n; \mu)$ reduces to the central t distribution, denoted as $t(n)$, which has the density

$$f(x) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\Gamma(1/2)n^{1/2}} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}.$$

Note that a $t(n)$ random variable is symmetric about zero, and its k^{th} moment exists only for $k < n$; when $n > 2$, its mean is zero and variance is $n/(n-2)$.

As n tends to infinity, it can be seen that

$$\left(1 + \frac{x^2}{n}\right)^{-(n+1)/2} = \left[\left(1 + \frac{x^2}{n}\right)^{n/x^2}\right]^{-x^2/2} \left(1 + \frac{x^2}{n}\right)^{-1/2} \to \exp(-x^2/2).$$

Also note that $\Gamma(1/2) = \pi^{1/2}$ and that for large n,

$$\frac{\Gamma((n+1)/2)}{\Gamma(n/2)} \approx (n/2)^{1/2}.$$

Thus, when n tends to infinity, the density of $t(n)$ converges to

$$\frac{1}{\sqrt{2\pi}} \exp(-x^2/2),$$

the density of the standard normal random variable. When $n = 1$, the density for $t(1)$ becomes

$$f(x) = \frac{1}{\pi[1 + x^2]}.$$

This is also the density of the *Cauchy* random variable with the location parameter 0. Note that the Cauchy random variable does not even have the first moment.

Let z_1 and z_2 be two independent random variables such that $z_1 \sim \chi^2(n_1; \nu_1)$ and $z_2 \sim \chi^2(n_2; \nu_2)$. Then,

$$\frac{z_1/n_1}{z_2/n_2} \sim F(n_1, n_2; \nu_1, \nu_2),$$

the non-central F distribution with the degrees of freedom n_1 and n_2 and the non-centrality parameters ν_1 and ν_2. The k^{th} moment of $F(n_1, n_2; \nu_1, \nu_2)$ exists when $k < n_2/2$. In many statistical applications we usually encounter $F(n_1, n_2; \nu_1, 0)$. When $n_2 > 2$, the mean of $F(n_1, n_2; \nu_1, 0)$is

$$\frac{n_2(n_1 + \nu_1)}{n_1(n_2 - 2)};$$

when $n_2 > 4$, the variance is

$$2\Big(\frac{n_2}{n_1}\Big)^2 \frac{(n_1+\nu_1)^2 + (n_1+2\nu_1)(n_2-2)}{(n_2-2)^2(n_2-4)}.$$

If both ν_1 and ν_2 are zero, we have the central F distribution $F(n_1, n_2)$. When $n_2 > 2$, $F(n_1, n_2)$ has mean $n_2/(n_2-2)$; when $n_2 > 4$, it has variance

$$\frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}.$$

Note that if a random variable is distributed as $t(n)$, its square has the $F(1, n)$ distribution.

2.4 Likelihood

Suppose that we postulate p as the joint probability function of the discrete random variables $z_1, \ldots, z_T$ with the parameter vector $\boldsymbol{\theta}$. Plugging the observed values $\zeta_1, \ldots, \zeta_T$ into p we then obtain a function of $\boldsymbol{\theta}$:

$$L(\boldsymbol{\theta}) := p(\zeta_1, \ldots, \zeta_T; \boldsymbol{\theta}).$$

This function represents the probability (likelihood) that those observed values are generated from the postulated probability function p; different parameter values of course result in different probability values. Thus, $L(\boldsymbol{\theta})$ is also known as the *likelihood function* of $\boldsymbol{\theta}$.

Similarly, let f denote the postulated joint density function of the random vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T$ with the parameter vector $\boldsymbol{\theta}$. Then given the observed values $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T$, the likelihood function of $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}) := f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta}).$$

In what follows, we will use L and f interchangeably. Note, however, that a postulated density function need not be the true density function that generates the random variables.

When f is differentiable and non-zero with probability one, the gradient vector of $\log L(\boldsymbol{\theta})$,

$$\nabla_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}) = \frac{1}{L(\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}),$$

is known as the *score* vector, denoted as $\boldsymbol{s}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})$. We can then write

$$\boldsymbol{s}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta}) f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta}).$$

For a given $\boldsymbol{\theta}$, the score vector may vary with the observed values $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T$. Thus, we can also treat the score vector as a random vector and denote it as $\boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta})$.

When differentiation and integration can be interchanged,

$$\begin{aligned}
&\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \boldsymbol{s}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\, f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\zeta}_1 \ \ldots\ \mathrm{d}\boldsymbol{\zeta}_T \\
&\quad = \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\zeta}_1 \ \ldots\ \mathrm{d}\boldsymbol{\zeta}_T \\
&\quad = \nabla_{\boldsymbol{\theta}} \left(\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\zeta}_1 \ \ldots\ \mathrm{d}\boldsymbol{\zeta}_T \right) \\
&\quad = \nabla_{\boldsymbol{\theta}}\, 1 \\
&\quad = \mathbf{0}.
\end{aligned}$$

The left-hand side is in effect the expectation of the score vector with respect to f. If there exists $\boldsymbol{\theta}_o$ such that $f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta}_o)$ is the true density function, we immediately obtain the following result.

Lemma 2.8 *If there exists $\boldsymbol{\theta}_o$ such that $f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta}_o)$ is the joint density function of the random vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T$. Then under regularity conditions,*

$$\mathbb{E}[\boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta}_o)] = \mathbf{0},$$

where $\boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta}_o)$ is the score evaluated at $\boldsymbol{\theta}_o$, and $\mathbb{E}$ is taken with respect to the true density function.

Remark: Lemma 2.8 requires the conditions that ensure differentiability of the likelihood function and interchangeability of differentiation and integration. We do not give those conditions explicitly; see e.g., Amemiya (1985) for some sufficient conditions. This comment also applies to Lemma 2.9.

It is easy to see that the Hessian matrix of the log-likelihood function is

$$\nabla^2_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}) = \frac{1}{L(\boldsymbol{\theta})} \nabla^2_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) - \frac{1}{L(\boldsymbol{\theta})^2} [\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})][\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})]',$$

where the second term is just the outer product of the score vector, and

$$\begin{aligned}
&\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \frac{1}{L(\boldsymbol{\theta})} \nabla^2_{\boldsymbol{\theta}} L(\boldsymbol{\theta})\, f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\zeta}_1 \cdots \mathrm{d}\boldsymbol{\zeta}_T \\
&\quad = \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \nabla^2_{\boldsymbol{\theta}} f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\zeta}_1 \cdots \mathrm{d}\boldsymbol{\zeta}_T \\
&\quad = \nabla^2_{\boldsymbol{\theta}} \left(\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\zeta}_1 \cdots \mathrm{d}\boldsymbol{\zeta}_T \right) \\
&\quad = \nabla^2_{\boldsymbol{\theta}} 1 \\
&\quad = \mathbf{0}.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \nabla^2_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta})\, f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\zeta}_1 \cdots \mathrm{d}\boldsymbol{\zeta}_T \\
&\quad = -\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \boldsymbol{s}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta}) \boldsymbol{s}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})' f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\zeta}_1 \cdots \mathrm{d}\boldsymbol{\zeta}_T,
\end{aligned}$$

where the left-hand side is the expectation of the Hessian matrix and the right-hand side is negative of the variance-covariance matrix of $\boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta})$, both with respect to the postulated density function f. If $f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta}_o)$ is the true density function, the variance-covariance matrix of $\boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta}_o)$ is known as the *information matrix*. Together with Lemma 2.8, we have the so-called *information matrix equality*.

Lemma 2.9 *If there exists $\boldsymbol{\theta}_o$ such that $f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta}_o)$ is the joint density function of the random vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T$. Then under regularity conditions,*

$$\mathbb{E}[\nabla^2_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}_o)] + \mathrm{var}(\boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta}_o)) = \mathbf{0},$$

where $\nabla^2_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}_o)$ is the Hessian matrix of $\log L$ evaluated at $\boldsymbol{\theta}_o$, and $\mathbb{E}$ and var *are taken with respect to the true density function.*

Remark: When f is not the true density function, Lemma 2.8 and 2.9 need not hold. That is, neither $\mathbb{E}[\boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta})]$ nor

$$\mathbb{E}[\nabla^2_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta})] + \mathrm{var}(\boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta}))$$

is necessarily zero.

2.5 Estimation

2.5.1 Point Estimation

Let $\boldsymbol{\theta}_o$ denote a parameter vector associated with the joint distribution of T random vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T$. A *point estimator* (or simply an estimator) for $\boldsymbol{\theta}_o$ is a function of these random vectors:

$$\hat{\boldsymbol{\theta}} = h(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T),$$

where h is some function. An estimator is clearly a random vector. Once the observed values of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T$ are plugged into this function, we obtain a *point estimate*. That is, a point estimate is just a particular value that an estimator may assume.

A simple principle of constructing estimators for moments is known as *analog estimation*. This principle suggests to estimate population moments using their finite-sample counterparts. For example, given a sample of T random variables $z_1, \ldots, z_T$ with the common k^{th} moment $\mathrm{I\!E}(z_1^k)$, the analog estimator for $\mathrm{I\!E}(z_1^k)$ is simply the sample average of z_i^k:

$$\frac{1}{T}\sum_{t=1}^{T} z_i^k.$$

In particular, the sample mean $\bar{z}$ is the analog estimator for the population mean.

To estimate the parameter vector $\boldsymbol{\theta}_o$, it is also natural to maximize the associated likelihood function $L(\boldsymbol{\theta})$. The resulting solution is known as the *maximum likelihood estimator* (MLE) for $\boldsymbol{\theta}_o$, denoted as $\tilde{\boldsymbol{\theta}}$ or $\tilde{\boldsymbol{\theta}}_T$, where the subscript T indicates that this is an estimator based on a sample of T observations. As the maximum of a function is invariant with respect to monotonic transformations, it is quite common to compute the MLE by maximizing the log-likelihood function $\log L(\boldsymbol{\theta})$. It follows that the score vector evaluated at $\tilde{\boldsymbol{\theta}}$ must be zero; i.e., $\boldsymbol{s}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \tilde{\boldsymbol{\theta}}) = \mathbf{0}$.

2.5.2 Criteria for Point Estimators

Let $\hat{\boldsymbol{\theta}}$ be an estimator for $\boldsymbol{\theta}_o$. The difference $\mathrm{I\!E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}_o$ is called the *bias* of $\hat{\boldsymbol{\theta}}$. An estimator is said to be *unbiased* if it has zero bias, i.e.,

$$\mathrm{I\!E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}_o;$$

otherwise, it is *biased*. Unbiasedness does not ensure that an estimate is close to the true parameter, however. In fact, it is even possible that all possible values of an unbiased estimator deviate from the true parameter by a constant.

Given two unbiased estimators, it is therefore natural to choose the one whose values are more concentrated around the true parameter. For real-valued unbiased estimators, this amounts to selecting an estimator with a smaller variance. If they are vector-valued, we adopt the following *efficiency* criterion. An unbiased estimator $\hat{\boldsymbol{\theta}}_1$ is said to be "better" (more efficient) than an unbiased estimator $\hat{\boldsymbol{\theta}}_2$ if

$$\mathrm{var}(\boldsymbol{a}'\hat{\boldsymbol{\theta}}_2) \geq \mathrm{var}(\boldsymbol{a}'\hat{\boldsymbol{\theta}}_1),$$

for all non-zero vectors $\boldsymbol{a}$. This is equivalent to the condition that

$$\boldsymbol{a}'[\mathrm{var}(\hat{\boldsymbol{\theta}}_2) - \mathrm{var}(\hat{\boldsymbol{\theta}}_1)]\boldsymbol{a} \geq 0,$$

i.e., $\mathrm{var}(\hat{\boldsymbol{\theta}}_2) - \mathrm{var}(\hat{\boldsymbol{\theta}}_1)$ is a positive semi-definite matrix. Given a class of unbiased estimators, if one of them is better than all other estimators in that class, it is the "best" (most efficient) within this class.

More generally, we can compare estimators based on mean squared error (MSE):

$$\begin{aligned}
&\mathbb{E}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)'] \\
&\quad = \mathbb{E}[(\hat{\boldsymbol{\theta}} - \mathbb{E}(\hat{\boldsymbol{\theta}}) + \mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}_o)(\hat{\boldsymbol{\theta}} - \mathbb{E}(\hat{\boldsymbol{\theta}}) + \mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}_o)'] \\
&\quad = \mathrm{var}(\hat{\boldsymbol{\theta}}) + [\mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}_o]\,[\mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}_o]',
\end{aligned}$$

where the second term is the outer product of the bias vector. An estimator $\hat{\boldsymbol{\theta}}_1$ (not necessarily unbiased) is said to be better (more efficient) than $\hat{\boldsymbol{\theta}}_2$ if $\mathrm{MSE}(\hat{\boldsymbol{\theta}}_2) - \mathrm{MSE}(\hat{\boldsymbol{\theta}}_1)$ is a positive semi-definite matrix. Clearly, the MSE criterion reduces to the previous variance-based criterion when estimators are unbiased.

The following result shows that the inverse of the information matrix is a lower bound, also known as the *Cramér-Rao lower bound*, for the variance-covariance matrix of any unbiased estimator.

Lemma 2.10 (Cramér-Rao) *If there exists $\boldsymbol{\theta}_o$ such that $f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta}_o)$ is the joint density function of the random vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T$. Let $\hat{\boldsymbol{\theta}}$ denote an unbiased estimator for $\boldsymbol{\theta}$ based on these random vectors. If $\mathrm{var}(\boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta}_o))$ is positive definite,*

$$\mathrm{var}(\hat{\boldsymbol{\theta}}) - \mathrm{var}(\boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta}_o))^{-1}$$

is a positive semi-definite matrix.

Proof: We first note that for any unbiased estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$,

$$\begin{aligned}
&\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\, \boldsymbol{s}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\, f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\zeta}_1 \cdots \mathrm{d}\boldsymbol{\zeta}_T \\
&\quad = \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \hat{\boldsymbol{\theta}}\, \boldsymbol{s}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\, f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\zeta}_1 \cdots \mathrm{d}\boldsymbol{\zeta}_T \\
&\quad = \nabla_{\boldsymbol{\theta}} \left(\int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} \hat{\boldsymbol{\theta}}\, f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\zeta}_1 \cdots \mathrm{d}\boldsymbol{\zeta}_T \right) \\
&\quad = \nabla_{\boldsymbol{\theta}} \boldsymbol{\theta} \\
&\quad = \boldsymbol{I},
\end{aligned}$$

where the third equality holds because $\hat{\boldsymbol{\theta}}$ is unbiased for $\boldsymbol{\theta}$ when $f(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{\theta})$ is the associated density function. Thus,

$$\mathrm{cov}(\hat{\theta}, \boldsymbol{s}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{\theta}_o)) = \boldsymbol{I}.$$

The assertion now follows from Lemma 2.5, the multivariate version of the Cauchy-Schwarz inequality. □

By Lemma 2.10, an unbiased estimator is the best if its variance-covariance matrix achieves the Cramér-Rao lower bound; this is not a necessary condition, however.

2.5.3 Interval Estimation

While a point estimate is a particular value representing the unknown parameter, *interval estimation* results in a range of values that may contain the unknown parameter with certain probability.

Suppose that there is an estimate $\hat{\theta}$ for the true parameter θ_o and a function $q(\hat{\theta}, \theta_o)$ whose distribution is known. Then, given a probability value γ, we can find suitable values a and b such that

$$\mathbb{P}\{a < q(\hat{\theta}, \theta_o) < b\} = \gamma.$$

Solving the inequality above for θ_o we may obtain an interval containing θ_o. This leads to the probability statement:

$$\mathbb{P}\{\alpha < \theta_o < \beta\} = \gamma,$$

where α and β depend on a, b, and $\hat{\theta}$. We can then conclude that we are $\gamma \times 100$ percent sure that the interval (α, β) contains θ_o. Here, γ is the *confidence coefficient*, and (α, β)

is the associated *confidence interval* for θ_o. Given the estimate $\hat{\theta}$, it is easily seen that the larger the value of γ, the wider is the associated confidence interval.

Let A_1 denote the event that a confidence interval contains θ_1 and A_2 the event that a confidence interval contains θ_2. The intersection $A = A_1 \cap A_2$ is thus the event that a confidence "box" covers both parameters. When A_1 and A_2 are independent such that $\mathbb{P}(A_1) = \mathbb{P}(A_2) = \gamma$, we have $\mathbb{P}(A) = \gamma^2$. When these two events are not independent (e.g., the parameter estimators of θ_1 and θ_2 are correlated), it becomes difficult to determine $\mathbb{P}(A)$. As such, finding a proper confidence "box" based on individual confidence intervals is by no means an easy job. On the other hand, if a function $q(\hat{\theta}_1, \hat{\theta}_2, \theta_1, \theta_2)$ with a known distribution is available, we can, for a given γ, find the values a and b such that

$$\mathbb{P}\{a < q(\hat{\theta}_1, \hat{\theta}_2, \theta_1, \theta_2) < b\} = \gamma.$$

By solving the inequality above for θ_1 and θ_2 we may obtain a *confidence region* in which the point (θ_1, θ_2) lies with probability γ.

2.6 Hypothesis Testing

2.6.1 Basic Concepts

Given a sample of data, it is often desirable to check if certain characteristics of the underlying random mechanism (population) are supported by these data. For this purpose, a *hypothesis* of these characteristics must be specified, and a *test* is constructed so as to generate a rule of rejecting or accepting (not rejecting) the postulated hypothesis.

The hypothesis being tested is called the *null hypothesis*, denoted as H_0; the other states or values of the characteristics of interest form an *alternative hypothesis*, denoted as H_1. Hypotheses are usually formulated in terms of the parameters of models. For example, one may specify that $H_0 : \boldsymbol{\theta}_o = \boldsymbol{a}$ for some $\boldsymbol{a}$ and $H_1 : \boldsymbol{\theta}_o \neq \boldsymbol{a}$. Here, H_0 is a *simple hypothesis* in the sense that the parameter vector being tested takes a single value, but H_1 is a *composite hypothesis* in that the parameter vector may take more than one values. Given a sample of random variables $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T$, a *test statistic* is a function of these random variables, denoted as $\mathcal{T}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T)$. The *critical region* C of $\mathcal{T}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T)$ is the range of its possible values that lead to rejection of the null hypothesis. In what follows, the set

$$\Gamma = \{\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T : \mathcal{T}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T) \in C\}$$

will also be referred to as the critical region of $\mathcal{T}$. The complement of the critical region, C^c, is the region containing the values of $\mathcal{T}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T)$ that lead to acceptance of the null hypothesis. We can also define

$$\Gamma^c = \{\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T : \mathcal{T}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T) \in C^c\}$$

as the acceptance region of $\mathcal{T}$.

A test may yield incorrect inferences. A test is said to commit the *type I error* if it rejects the null hypothesis when the null hypothesis is in fact true; a test is said to commit the *type II error* if it accepts the null hypothesis when the alternative hypothesis is true. Suppose that we are interested in testing $H_0 : \boldsymbol{\theta}_o = \boldsymbol{a}$ against $H_1 : \boldsymbol{\theta}_o = \boldsymbol{b}$. Let $\mathbb{P}_0$ be the probability when $\boldsymbol{\theta}_o = \boldsymbol{a}$ and $\mathbb{P}_1$ the probability when $\boldsymbol{\theta}_o = \boldsymbol{b}$. The probability of the type I error is then

$$\alpha = \mathbb{P}_0((\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T) \in \Gamma) = \int_\Gamma f_0(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{a})\, \mathrm{d}\boldsymbol{\zeta}_1 \cdots \mathrm{d}\boldsymbol{\zeta}_T,$$

where $f_0(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{a})$ is the joint density with the parameter $\boldsymbol{\theta}_o = \boldsymbol{a}$. The value α is also known as the *size* or *significance level* of the test. The probability of the type II error is

$$\beta = \mathbb{P}_1((\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T) \in \Gamma^c) = \int_{\Gamma^c} f_1(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T; \boldsymbol{b})\, \mathrm{d}\boldsymbol{\zeta}_1 \cdots \mathrm{d}\boldsymbol{\zeta}_T,$$

where $f_1(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T; \boldsymbol{b})$ is the joint density with the parameter $\boldsymbol{\theta}_o = \boldsymbol{b}$. Clearly, α decreases when the critical region Γ is smaller; in the mean time, β increases due to a larger Γ^c. Thus, there is usually a trade-off between these two error probabilities.

Note, however, that the probability of the type II error cannot be defined as above when the alternative hypothesis is composite: $\boldsymbol{\theta}_o \in \Theta_1$, where Θ_1 is a set of parameter values in the parameter space. Consider now the probability $1 - \mathbb{P}_1(\Gamma^c) = \mathbb{P}_1(\Gamma)$, which is the probability of rejecting the null hypothesis when H_1 is true. Thus, both $\mathbb{P}_0(\Gamma)$ and $\mathbb{P}_1(\Gamma)$ are the probabilities of rejecting the null hypothesis under two different parameter values. More generally, define the *power function* of the test as

$$\pi(\boldsymbol{\theta}_o) = \mathbb{P}_{\boldsymbol{\theta}_o}\{(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T) \in \Gamma\},$$

where $\boldsymbol{\theta}_o$ varies in the parameter space. In particular, $\pi(\boldsymbol{a}) = \alpha$. For $\boldsymbol{\theta}_o \in \Theta_1$, $\pi(\boldsymbol{\theta}_o)$ describes the ability of a test that can correctly detect the falsity of the null hypothesis; these probabilities are also referred to as the *powers* of the test. The probability of the type II error under the composite alternative hypothesis $\boldsymbol{\theta}_o \in \Theta_1$ can now be defined as

$$\beta = \max_{\boldsymbol{\theta}_o \in \Theta_1} [1 - \pi(\boldsymbol{\theta}_o)].$$

2.6.2 Construction of Tests

Given the null hypothesis $\boldsymbol{\theta}_o = \boldsymbol{a}$, the test statistic $\mathcal{T}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T)$ is usually based on the comparison of an estimator of $\boldsymbol{\theta}_o$ and the hypothesized value $\boldsymbol{a}$. This statistic must have a known distribution under the null hypothesis, which will be referred to as the *null distribution*.

Given the statistic $\mathcal{T}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T)$, the probability $\mathbb{P}_0(\mathcal{T}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T) \in C)$ can be determined by the null distribution of $\mathcal{T}$. If this probability is small, the event that $\mathcal{T}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T) \in C$ would be considered "unlikely" or "improbable" under the null hypothesis, while the event that $\mathcal{T}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T) \in C^c$ would be considered "likely" or "probable". If the former event does occur (i.e., for data $\boldsymbol{z}_1 = \boldsymbol{\zeta}_1, \ldots, \boldsymbol{z}_T = \boldsymbol{\zeta}_T$, $\mathcal{T}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T)$ falls in C), it constitutes an evidence against the null hypothesis, so that the null hypothesis is rejected; otherwise, we accept (do not reject) the null hypothesis. Therefore, one should specify a small significance level α and determine the associated critical region C by

$$\alpha = \mathbb{P}_0\{\mathcal{T}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T) \in C\}.$$

As such, we shall write the critical region for the significance level α as C_α. This approach ensures that, even though the decision of rejection might be wrong, the probability of making the type I error is no greater than α. A test statistic is said to be *significant* if it is in the critical region; otherwise, it is *insignificant*.

Another approach is to reject the null hypothesis if

$$\mathbb{P}_0\{v : v > \mathcal{T}(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_T)\}$$

is small. This probability is the tail probability of the null distribution and also known as the p-value of the statistic $\mathcal{T}$. Although this approach does not require specifying the critical region, it is virtually the same as the previous approach.

The rationale of our test decision is that the null hypothesis is rejected because the test statistic takes an unlikely value. It is then natural to expect that the calculated statistic is relatively more likely under the alternative hypothesis. Given the null hypothesis $\boldsymbol{\theta}_o = \boldsymbol{a}$ and alternative hypothesis $\boldsymbol{\theta}_o \in \Theta_1$, we would like to have a test such that

$$\pi(\boldsymbol{a}) \leq \pi(\boldsymbol{\theta}_o), \qquad \boldsymbol{\theta}_o \in \Theta_1.$$

A test is said to be *unbiased* if its size is no greater than the powers under the alternative hypothesis. Moreover, we would like to have a test that can detect the falsity of the

null hypothesis with probability approaching one when there is sufficient information. That is, for every $\boldsymbol{\theta}_o \in \Theta_1$,

$$\pi(\boldsymbol{\theta}_o) = \mathbb{P}_{\boldsymbol{\theta}_o}\{\mathcal{T}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T) \in C\} \to 1,$$

as $T \to \infty$. A test is said to be *consistent* if its power approaches one when the sample size becomes infinitely large.

Example 2.11 Given the sample of i.i.d. normal random variables $z_1, \ldots, z_T$ with mean μ_o and variance one. We would like to test the null hypothesis $\mu_o = 0$. A natural estimator for μ_o is the sample average $\bar{z} = T^{-1}\sum_{t=1}^{T} z_t$. It is well known that

$$\sqrt{T}(\bar{z} - \mu_o) \sim N(0, 1).$$

Hence, $\sqrt{T}\bar{z} \sim N(0, 1)$ under the null hypothesis; that is, the null distribution of the statistic $\sqrt{T}\bar{z}$ is the standard normal distribution. Given the significance level α, we can determine the critical region C_α using

$$\alpha = \mathbb{P}_0(\sqrt{T}\bar{z} \in C_\alpha).$$

Let Φ denote the distribution function of the standard normal random variable. For $\alpha = 0.05$, we know

$$0.05 = \mathbb{P}_0(\sqrt{T}\bar{z} > 1.645) = 1 - \Phi(1.645).$$

The critical region is then $(1.645, \infty)$; the null hypothesis is rejected if the calculated statistic falls in this interval. When the null hypothesis is false, the distribution of $\sqrt{T}\bar{z}$ is no longer $N(0, 1)$ but is $N(\mu_o, 1)$. Suppose that $\mu_o > 0$. Then,

$$\mathbb{P}_1(\sqrt{T}\bar{z} > 1.645) = \mathbb{P}_1(\sqrt{T}(\bar{z} - \mu_o) > 1.645 - \sqrt{T}\mu_o).$$

Since $\sqrt{T}(\bar{z} - \mu_o) \sim N(0, 1)$ under the alternative hypothesis, we have the power:

$$\mathbb{P}_1(\sqrt{T}\bar{z} > 1.645) = 1 - \Phi(1.645 - \sqrt{T}\mu_o).$$

Given that $\mu_o > 0$, this probability must be greater than the test size (0.05), so that the test is unbiased. On the other hand, when T increases, $1.645 - \sqrt{T}\mu_o$ becomes even smaller, so that the power improves. When T tends to infinity, the power approaches one, so that $\sqrt{T}\bar{z}$ is a consistent test. □

References

Amemiya, Takeshi (1985). *Advanced Econometrics*, Cambridge, MA: Harvard University Press.

Amemiya, Takeshi (1994). *Introduction to Statistics and Econometrics*, Cambridge, MA: Harvard University Press.

Rudin, Walter (1976). *Principles of Mathematical Analysis*, Third edition, New York, NY: McGraw-Hill.

Chapter 3

Classical Least Squares Theory

3.1 Introduction

Economists have proposed numerous hypotheses and theories in order to describe the behavior of economic agents and the relationships between economic variables. Although these propositions may be theoretically appealing and logically correct, they need not be practically relevant unless they are supported by real world data. A theory with empirical evidence is of course more convincing. Therefore, empirical analysis has become an indispensable ingredient of contemporary economic research. By *econometrics* we mean the statistical and mathematical methods that can be used to analyze empirical relationships between economic variables.

A leading approach in econometrics is the *regression* analysis in which a regression model of a collection of explanatory variables is specified to characterize the behavior of the variable of interest. The simplest and most commonly used specification is the linear model. Once a linear model is specified, it remains to estimate unknown model parameters, test economic and econometric hypotheses, and draw inferences from these results. This chapter is concerned with the most important estimation method in linear regression, the method of *ordinary least squares*. Readers can also find related topics in many econometrics textbooks, e.g., Davidson and MacKinnon (1993), Goldberger (1991), Greene (2000), Harvey (1990), Intriligator et al. (1996), Johnston (1984), Judge et al. (1988), Maddala (1992), Ruud (2000), and Theil (1971), among others.

3.2 The Method of Ordinary Least Squares

Suppose that there is a variable, y, whose behavior over time (or across individual units) is of interest to us. A theory may suggest that the behavior of y can be well characterized by some function f of the variables $x_1, \ldots, x_k$. Then, $f(x_1, \ldots, x_k)$ may be viewed as a "systematic" component of y provided that no other variables can further account for the residual behavior, $y - f(x_1, \ldots, x_k)$. In the context of linear regression, the function f is specified as a linear function. The method of *ordinary least squares* (OLS) enables us to determine the linear weights (parameters) of this specification.

3.2.1 Simple Linear Regression

In simple linear regression, only one variable x is designated to describe the behavior of the variable y. The linear specification is

$$\alpha + \beta x,$$

where α and β are unknown parameters. We can then write

$$y = \alpha + \beta x + e(\alpha, \beta),$$

where $e(\alpha, \beta) = y - \alpha - \beta x$ denotes the error resulted from this specification. In what follows, y will be referred to as the *dependent variable* (*regressand*), and x will be referred to as an *explanatory variable* (*regressor*). Note that the regressor x itself may be a function of some other variables, e.g., $x = z^2$ or $x = \log z$.

Suppose that we have T observations of the variables y and x. Given the linear specification above, our objective is to find suitable α and β such that the resulting linear function "best" fits the data (y_t, x_t), $t = 1, \ldots, T$. Here, the generic subscript t is used for both cross-section and time-series data. The OLS method suggests to find a straight line whose sum of squared errors is as small as possible. This amounts to find α and β that minimize the following OLS criterion function:

$$Q(\alpha, \beta) := \frac{1}{T} \sum_{t=1}^{T} e_t(\alpha, \beta)^2 = \frac{1}{T} \sum_{t=1}^{T} (y_t - \alpha - \beta x_t)^2.$$

The solutions can be easily obtained by solving the first order conditions.

The first order conditions of this minimization problem are:

$$\frac{\partial Q(\alpha,\beta)}{\partial \alpha} = -\frac{2}{T}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t) = 0,$$

$$\frac{\partial Q(\alpha,\beta)}{\partial \beta} = -\frac{2}{T}\sum_{t=1}^{T}(y_t - \alpha - \beta x_t)x_t = 0.$$

Solving for α and β we have the following solutions:

$$\hat{\beta}_T = \frac{\sum_{t=1}^{T}(y_t - \bar{y})(x_t - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2},$$

$$\hat{\alpha}_T = \bar{y} - \hat{\beta}_T \bar{x},$$

where $\bar{y} = \sum_{t=1}^{T} y_t/T$ and $\bar{x} = \sum_{t=1}^{T} x_t/T$. As $\hat{\alpha}_T$ and $\hat{\beta}_T$ are obtained by minimizing the OLS criterion function, they are known as the OLS estimators of α and β, respectively. The subscript T of $\hat{\alpha}_T$ and $\hat{\beta}_T$ signifies that these solutions are obtained from a sample of T observations. Note that if x_t is a constant c for every t, then $\bar{x} = c$, and hence $\hat{\beta}_T$ cannot be computed.

The function $\hat{y} = \hat{\alpha}_T + \hat{\beta}_T x$ is the estimated *regression line* with the intercept $\hat{\alpha}_T$ and slope $\hat{\beta}_T$. We also say that this line is obtained by regressing y on (the constant one and) the regressor x. The regression line so computed gives the "best" fit of data, in the sense that any other linear function of x would yield a larger sum of squared errors. For a given x_t, the OLS fitted value is a point on the regression line:

$$\hat{y}_t = \hat{\alpha}_T + \hat{\beta}_T x_t.$$

The difference between y_t and $\hat{y}_t$ is the t^{th} OLS *residual*:

$$\hat{e}_t := y_t - \hat{y}_t,$$

which corresponds to the error of the specification as

$$\hat{e}_t = e_t(\hat{\alpha}_T, \hat{\beta}_T).$$

Note that regressing y on x and regressing x on y lead to different regression lines in general, except when all (y_t, x_t) lie on the same line; see Exercise 3.9.

Remark: Different criterion functions would result in other estimators. For example, the so-called *least absolute deviation* estimator can be obtained by minimizing the average of the sum of absolute errors:

$$\frac{1}{T}\sum_{t=1}^{T}|y_t - \alpha - \beta x_t|,$$

which in turn determines a different regression line. We refer to Manski (1991) for a comprehensive discussion of this topic.

3.2.2 Multiple Linear Regression

More generally, we may specify a linear function with k explanatory variables to describe the behavior of y:

$$\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k,$$

so that

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e(\beta_1, \ldots, \beta_k),$$

where $e(\beta_1, \ldots, \beta_k)$ again denotes the error of this specification. Given a sample of T observations, this specification can also be expressed as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}(\boldsymbol{\beta}), \tag{3.1}$$

where $\boldsymbol{\beta} = (\beta_1 \;\; \beta_2 \;\; \cdots \;\; \beta_k)'$ is the vector of unknown parameters, $\boldsymbol{y}$ and $\boldsymbol{X}$ contain all the observations of the dependent and explanatory variables, i.e.,

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T1} & x_{T2} & \cdots & x_{Tk} \end{bmatrix},$$

where each column of $\boldsymbol{X}$ contains T observations of an explanatory variable, and $\boldsymbol{e}(\boldsymbol{\beta})$ is the vector of errors. It is typical to set the first explanatory variable as the constant one so that the first column of $\boldsymbol{X}$ is the $T \times 1$ vector of ones, $\boldsymbol{\ell}$. For convenience, we also write $\boldsymbol{e}(\boldsymbol{\beta})$ as $\boldsymbol{e}$ and its element $e_t(\boldsymbol{\beta})$ as e_t.

Our objective now is to find a k-dimensional regression hyperplane that "best" fits the data $(\boldsymbol{y}, \boldsymbol{X})$. In the light of Section 3.2.1, we would like to minimize, with respect to $\boldsymbol{\beta}$, the average of the sum of squared errors:

$$Q(\boldsymbol{\beta}) := \frac{1}{T}\boldsymbol{e}(\boldsymbol{\beta})'\boldsymbol{e}(\boldsymbol{\beta}) = \frac{1}{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{3.2}$$

This is a well-defined problem provided that the basic *identification requirement* below holds for the specification (3.1).

[ID-1] The $T \times k$ data matrix $\boldsymbol{X}$ is of full column rank k.

Under [ID-1], the number of regressors, k, must be strictly less than the number of observations, T. This is so because if $k > T$, the rank of $\boldsymbol{X}$ must be less than or equal to T, and hence $\boldsymbol{X}$ cannot have full column rank. Moreover, [ID-1] requires that any linear specification does not contain any "redundant" regressor; that is, any column vector of $\boldsymbol{X}$ cannot be written as a linear combination of other column vectors. For example, $\boldsymbol{X}$ contains a column of ones and a column of x_t in simple linear regression. These two columns would be linearly dependent if $x_t = c$ for every t. Thus, [ID-1] requires that x_t in simple linear regression is not a constant.

The first order condition of the OLS minimization problem is

$$\nabla_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} (\boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta})/T = \boldsymbol{0}.$$

By the matrix differentiation results in Section 1.2, we have

$$\nabla_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = -2\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})/T = \boldsymbol{0}.$$

Equivalently, we can write

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y}. \tag{3.3}$$

These k equations, also known as the *normal equations*, contain exactly k unknowns. Given [ID-1], $\boldsymbol{X}$ is of full column rank so that $\boldsymbol{X}'\boldsymbol{X}$ is positive definite and hence invertible by Lemma 1.13. It follows that the unique solution to the first order condition is

$$\hat{\boldsymbol{\beta}}_T = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{3.4}$$

Moreover, the second order condition is also satisfied because

$$\nabla^2_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}) = 2(\boldsymbol{X}'\boldsymbol{X})/T$$

is a positive definite matrix under [ID-1]. Thus, $\hat{\boldsymbol{\beta}}_T$ is the unique minimizer of the OLS criterion function and hence known as the OLS estimator of $\boldsymbol{\beta}$. This result is formally stated below.

Theorem 3.1 *Given the specification* (3.1)*, suppose that* [ID-1] *holds. Then, the OLS estimator* $\hat{\boldsymbol{\beta}}_T$ *given by* (3.4) *uniquely minimizes the OLS criterion function* (3.2)*.*

If $\boldsymbol{X}$ is not of full column rank, its column vectors are linearly dependent and therefore satisfy an exact linear relationship. This is the problem of *exact multicollinearity*.

In this case, $\boldsymbol{X}'\boldsymbol{X}$ is not invertible so that there exist infinitely many solutions to the normal equations $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y}$. As such, the OLS estimator $\hat{\boldsymbol{\beta}}_T$ cannot be uniquely determined. See Exercise 3.4 for a geometric interpretation of this result. Exact multicollinearity usually arises from inappropriate model specifications. For example, including both total income, total wage income, and total non-wage income as regressors results in exact multicollinearity because total income is, by definition, the sum of wage and non-wage income. See also Section 3.6.2 for another example. In what follows, the identification requirement for the linear specification (3.1) is always assumed.

Remarks:

1. Theorem 3.1 does not depend on the "true" relationship between $\boldsymbol{y}$ and $\boldsymbol{X}$. That is, whether (3.1) is a correct specification is irrelevant to the existence and uniqueness of the OLS estimator.

2. It is easy to verify that the magnitudes of the coefficient estimates $\hat{\beta}_i$, $i = 1, \ldots, k$, are affected by the measurement units of dependent and explanatory variables; see Exercise 3.7. As such, a larger coefficient estimate does not necessarily imply that the associated regressor is more important in explaining the behavior of $\boldsymbol{y}$. In fact, the coefficient estimates are not directly comparable in general; cf. Exercise 3.5.

Once the OLS estimator $\hat{\boldsymbol{\beta}}_T$ is obtained, we can plug it into the original linear specification and obtain the vector of OLS fitted values:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_T.$$

The vector of OLS residuals is then

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{e}(\hat{\boldsymbol{\beta}}_T).$$

From the normal equations (3.3) we can deduce the following algebraic results. First, the OLS residual vector must satisfy the normal equations:

$$\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0},$$

so that $\boldsymbol{X}'\hat{\boldsymbol{e}} = \boldsymbol{0}$. When $\boldsymbol{X}$ contains a column of constants (i.e., a column of $\boldsymbol{X}$ is $c\boldsymbol{\ell}$, where $\boldsymbol{\ell}$ is the vector of ones), $\boldsymbol{X}'\hat{\boldsymbol{e}} = \boldsymbol{0}$ implies

$$\boldsymbol{\ell}'\hat{\boldsymbol{e}} = \sum_{t=1}^{T} \hat{e}_t = 0.$$

That is, the sum of OLS residuals must be zero. Second,

$$\hat{\boldsymbol{y}}'\hat{\boldsymbol{e}} = \hat{\boldsymbol{\beta}}_T'\boldsymbol{X}'\hat{\boldsymbol{e}} = 0.$$

These results are summarized below.

Theorem 3.2 *Given the specification* (3.1)*, suppose that* [ID-1] *holds. Then, the vector of OLS fitted values* $\hat{\boldsymbol{y}}$ *and the vector of OLS residuals* $\hat{\boldsymbol{e}}$ *have the following properties.*

(a) $\boldsymbol{X}'\hat{\boldsymbol{e}} = \boldsymbol{0}$*; in particular, if* $\boldsymbol{X}$ *contains a column of constants,* $\sum_{t=1}^{T}\hat{e}_t = 0$*.*

(b) $\hat{\boldsymbol{y}}'\hat{\boldsymbol{e}} = 0$*.*

Note that when $\boldsymbol{\ell}'\hat{\boldsymbol{e}} = \boldsymbol{\ell}'(\boldsymbol{y} - \hat{\boldsymbol{y}}) = 0$, we have

$$\frac{1}{T}\sum_{t=1}^{T} y_t = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_t.$$

That is, the sample average of the data y_t is the same as the sample average of the fitted values $\hat{y}_t$ when $\boldsymbol{X}$ contains a column of constants.

3.2.3 Geometric Interpretations

The OLS estimation result has nice geometric interpretations. These interpretations have nothing to do with the stochastic properties to be discussed in Section 3.3, and they are valid as long as the OLS estimator exists.

In what follows, we write $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ which is an orthogonal projection matrix that projects vectors onto span($\boldsymbol{X}$) by Lemma 1.14. The vector of OLS fitted values can be written as

$$\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y}.$$

Hence, $\hat{\boldsymbol{y}}$ is the orthogonal projection of $\boldsymbol{y}$ onto span($\boldsymbol{X}$). The OLS residual vector is

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{y},$$

which is the orthogonal projection of $\boldsymbol{y}$ onto span($\boldsymbol{X}$)$^{\perp}$ and hence is orthogonal to $\hat{\boldsymbol{y}}$ and $\boldsymbol{X}$; cf. Theorem 3.2. Consequently, $\hat{\boldsymbol{y}}$ is the "best approximation" of $\boldsymbol{y}$, given the information contained in $\boldsymbol{X}$, as shown in Lemma 1.10. Figure 3.1 illustrates a simple case where there are only two explanatory variables in the specification.

The following results are useful in many applications.

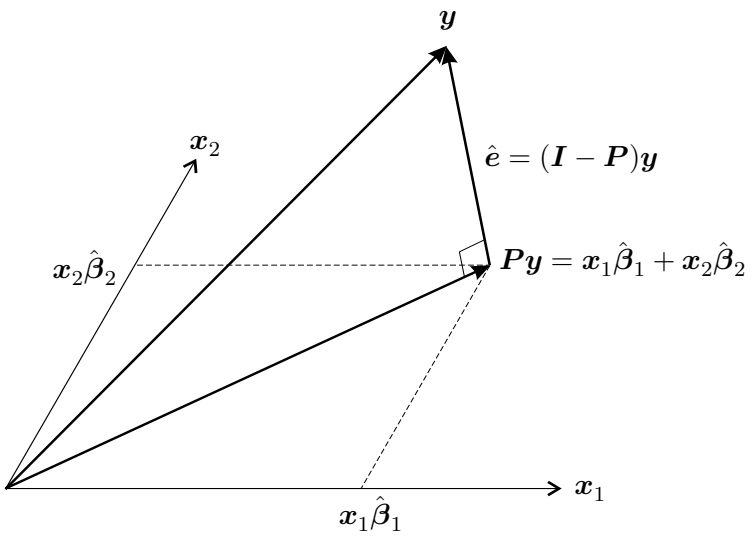

Figure 3.1: The orthogonal projection of $\boldsymbol{y}$ onto span($\boldsymbol{x}_1$,$\boldsymbol{x}_2$)

Theorem 3.3 (Frisch-Waugh-Lovell) *Given the specification*

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e},$$

where $\boldsymbol{X}_1$ is of full column rank k_1 and $\boldsymbol{X}_2$ is of full column rank k_2, let $\hat{\boldsymbol{\beta}}_T = (\hat{\boldsymbol{\beta}}_{1,T}' \; \hat{\boldsymbol{\beta}}_{2,T}')'$ denote the corresponding OLS estimators. Then,

$$\hat{\boldsymbol{\beta}}_{1,T} = [\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1]^{-1}\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{y},$$

$$\hat{\boldsymbol{\beta}}_{2,T} = [\boldsymbol{X}_2'(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{X}_2]^{-1}\boldsymbol{X}_2'(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y},$$

where $\boldsymbol{P}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$ and $\boldsymbol{P}_2 = \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'$.

Proof: These results can be directly verified from (3.4) using the matrix inversion formula in Section 1.4. Alternatively, write

$$\boldsymbol{y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_{1,T} + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_{2,T} + (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y},$$

where $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ with $\boldsymbol{X} = [\boldsymbol{X}_1 \; \boldsymbol{X}_2]$. Pre-multiplying both sides by $\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)$, we have

$$\begin{aligned}&\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{y} \\ &\quad = \boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_{1,T} + \boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_{2,T} + \boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}.\end{aligned}$$

The second term on the right-hand side vanishes because $(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_2 = \boldsymbol{0}$. For the third term, we know span($\boldsymbol{X}_2$) $\subseteq$ span($\boldsymbol{X}$), so that span($\boldsymbol{X}$)$^\perp \subseteq$ span($\boldsymbol{X}_2$)$^\perp$. As each

column vector of $\boldsymbol{I}-\boldsymbol{P}$ is in $\text{span}(\boldsymbol{X})^{\perp}$, $\boldsymbol{I}-\boldsymbol{P}$ is not affected if it is pojected onto $\text{span}(\boldsymbol{X}_2)^{\perp}$. That is,

$$(\boldsymbol{I}-\boldsymbol{P}_2)(\boldsymbol{I}-\boldsymbol{P})=\boldsymbol{I}-\boldsymbol{P}.$$

Similarly, $\boldsymbol{X}_1$ is in $\text{span}(\boldsymbol{X})$, and hence $(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{X}_1=\boldsymbol{0}$. It follows that

$$\boldsymbol{X}_1'(\boldsymbol{I}-\boldsymbol{P}_2)\boldsymbol{y}=\boldsymbol{X}_1'(\boldsymbol{I}-\boldsymbol{P}_2)\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_{1,T},$$

from which we obtain the expression for $\hat{\boldsymbol{\beta}}_{1,T}$. The proof for $\hat{\boldsymbol{\beta}}_{2,T}$ is similar. □

This result shows that $\hat{\boldsymbol{\beta}}_{1,T}$ can be computed from regressing $(\boldsymbol{I}-\boldsymbol{P}_2)\boldsymbol{y}$ on $(\boldsymbol{I}-\boldsymbol{P}_2)\boldsymbol{X}_1$, where $(\boldsymbol{I}-\boldsymbol{P}_2)\boldsymbol{y}$ and $(\boldsymbol{I}-\boldsymbol{P}_2)\boldsymbol{X}_1$ are the residual vectors of the "purging" regressions of $\boldsymbol{y}$ on $\boldsymbol{X}_2$ and $\boldsymbol{X}_1$ on $\boldsymbol{X}_2$, respectively. Similarly, $\hat{\boldsymbol{\beta}}_{2,T}$ can be obtained by regressing $(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{y}$ on $(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2$, where $(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{y}$ and $(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2$ are the residual vectors of the regressions of $\boldsymbol{y}$ on $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$, respectively.

From Theorem 3.3 we can deduce the following results. Consider the regression of $(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{y}$ on $(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2$. By Theorem 3.3 we have

$$(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{y}=(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_{2,T}+\text{ residual vector}, \tag{3.5}$$

where the residual vector is

$$(\boldsymbol{I}-\boldsymbol{P}_1)(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{y}=(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{y}.$$

Thus, the residual vector of (3.5) is identical to the residual vector of regressing $\boldsymbol{y}$ on $\boldsymbol{X}=[\boldsymbol{X}_1\ \boldsymbol{X}_2]$. Note that $(\boldsymbol{I}-\boldsymbol{P}_1)(\boldsymbol{I}-\boldsymbol{P})=\boldsymbol{I}-\boldsymbol{P}$ implies $\boldsymbol{P}_1=\boldsymbol{P}_1\boldsymbol{P}$. That is, the orthogonal projection of $\boldsymbol{y}$ directly on $\text{span}(\boldsymbol{X}_1)$ is equivalent to performing iterated projections of $\boldsymbol{y}$ on $\text{span}(\boldsymbol{X})$ and then on $\text{span}(\boldsymbol{X}_1)$. The orthogonal projection part of (3.5) now can be expressed as

$$(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_{2,T}=(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{P}\boldsymbol{y}=(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{y}.$$

These relationships are illustrated in Figure 3.2.

Similarly, we have

$$(\boldsymbol{I}-\boldsymbol{P}_2)\boldsymbol{y}=(\boldsymbol{I}-\boldsymbol{P}_2)\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_{1,T}+\text{ residual vector},$$

where the residual vector is also $(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{y}$, and the orthogonal projection part of this regression is $(\boldsymbol{P}-\boldsymbol{P}_2)\boldsymbol{y}$. See also Davidson and MacKinnon (1993) for more details.

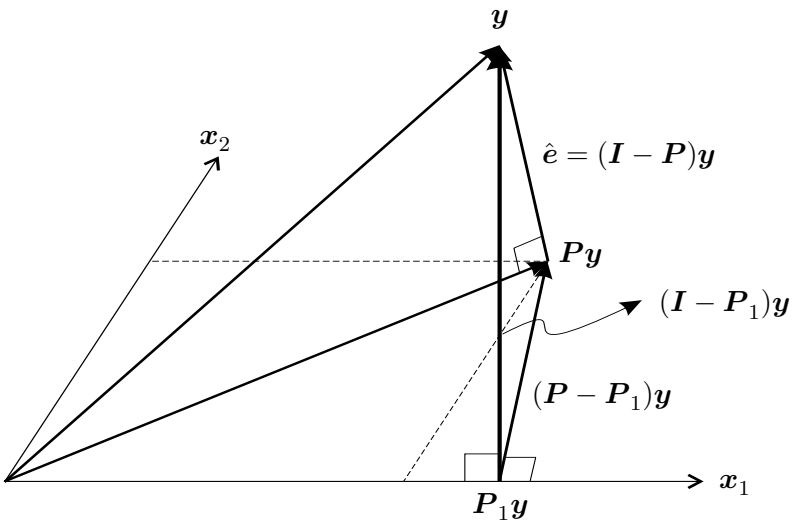

Figure 3.2: An illustration of the Frisch-Waugh-Lovell Theorem

Intuitively, Theorem 3.3 suggests that $\hat{\boldsymbol{\beta}}_{1,T}$ in effect describes how $\boldsymbol{X}_1$ characterizes $\boldsymbol{y}$, after the effect of $\boldsymbol{X}_2$ is excluded. Thus, $\hat{\boldsymbol{\beta}}_{1,T}$ is different from the OLS estimator of regressing $\boldsymbol{y}$ on $\boldsymbol{X}_1$ because the effect of $\boldsymbol{X}_2$ is not controlled in the latter. These two estimators would be the same if $\boldsymbol{P}_2\boldsymbol{X}_1 = \boldsymbol{0}$, i.e., $\boldsymbol{X}_1$ is orthogonal to $\boldsymbol{X}_2$. Also, $\hat{\boldsymbol{\beta}}_{2,T}$ describes how $\boldsymbol{X}_2$ characterizes $\boldsymbol{y}$, after the effect of $\boldsymbol{X}_1$ is excluded, and it is different from the OLS estimator from regressing $\boldsymbol{y}$ on $\boldsymbol{X}_2$, unless $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are orthogonal to each other.

As an application, consider the specification with $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$, where $\boldsymbol{X}_1$ contains the constant term and a time trend variable t, and $\boldsymbol{X}_2$ includes the other $k-2$ explanatory variables. This specification is useful when the variables of interest exhibit a trending behavior. Then, the OLS estimators of the coefficients of $\boldsymbol{X}_2$ are the same as those obtained from regressing (detrended) $\boldsymbol{y}$ on detrended $\boldsymbol{X}_2$, where detrended $\boldsymbol{y}$ and $\boldsymbol{X}_2$ are the residuals of regressing $\boldsymbol{y}$ and $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$, respectively. See Exercise 3.11 for another application.

3.2.4 Measures of Goodness of Fit

We have learned that from previous sections that, when the explanatory variables in a linear specification are given, the OLS method yields the best fit of data. In practice, one may consdier a linear specfication with different sets of regressors and try to choose a particular one from them. It is therefore of interest to compare the performance across

different specifications. In this section we discuss how to measure the *goodness of fit* of a specification. A natural goodness-of-fit measure is of course the sum of squared errors $\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$. Unfortunately, this measure is not invariant with respect to measurement units of the dependent variable and hence is not appropriate for model comparison. Instead, we consider the following "relative" measures of goodness of fit.

Recall from Theorem 3.2(b) that $\hat{\boldsymbol{y}}'\hat{\boldsymbol{e}} = 0$. Then,

$$\boldsymbol{y}'\boldsymbol{y} = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} + \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} + 2\hat{\boldsymbol{y}}'\hat{\boldsymbol{e}} = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} + \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}.$$

This equation can be written in terms of sum of squares:

$$\underbrace{\sum_{t=1}^{T} y_t^2}_{\text{TSS}} = \underbrace{\sum_{t=1}^{T} \hat{y}_t^2}_{\text{RSS}} + \underbrace{\sum_{t=1}^{T} \hat{e}_t^2}_{\text{ESS}},$$

where TSS stands for *total sum of squares* and is a measure of total squared variations of y_t, RSS stands for *regression sum of squares* and is a measures of squared variations of fitted values, and ESS stands for *error sum of squares* and is a measure of squared variation of residuals. The non-centered *coefficient of determination* (or non-centered R^2) is defined as the proportion of TSS that can be explained by the regression hyperplane:

$$R^2 = \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\text{ESS}}{\text{TSS}}. \tag{3.6}$$

Clearly, $0 \leq R^2 \leq 1$, and the larger the R^2, the better the model fits the data. In particular, a model has a perfect fit if $R^2 = 1$, and it does not account for any variation of $\boldsymbol{y}$ if $R^2 = 0$. It is also easy to verify that this measure does not depend on the measurement units of the dependent and explanatory variables; see Exercise 3.7.

As $\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} = \hat{\boldsymbol{y}}'\boldsymbol{y}$, we can also write

$$R^2 = \frac{\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}}{\boldsymbol{y}'\boldsymbol{y}} = \frac{(\hat{\boldsymbol{y}}'\boldsymbol{y})^2}{(\boldsymbol{y}'\boldsymbol{y})(\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}})}.$$

It follows from the discussion of inner product and Euclidean norm in Section 1.2 that the right-hand side is just $\cos^2\theta$, where θ is the angle between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$. Thus, R^2 can be interpreted as a measure of the linear association between these two vectors. A perfect fit is equivalent to the fact that $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ are collinear, so that $\boldsymbol{y}$ must be in $\text{span}(\boldsymbol{X})$. When $R^2 = 0$, $\boldsymbol{y}$ is orthogonal to $\hat{\boldsymbol{y}}$ so that $\boldsymbol{y}$ is in $\text{span}(\boldsymbol{X})^\perp$.

It can be verified that when a constant is added to all observations of the dependent variable, the resulting coefficient of determination also changes. This is clearly a

drawback because a sensible measure of fit should not be affected by the location of the dependent variable. Another drawback of the coefficient of determination is that it is non-decreasing in the number of variables in the specification. That is, adding more variables to a linear specification will *not* reduce its R^2. To see this, consider a specification with k_1 regressors and a more complex one containing the same k_1 regressors and additional k_2 regressors. In this case, the former specification is "nested" in the latter, in the sense that the former can be obtained from the latter by setting the coefficients of those additional regressors to zero. Since the OLS method searches for the best fit of data without any constraint, the more complex model cannot have a worse fit than the specifications nested in it. See also Exercise 3.8.

A measure that is invariant with respect to constant addition is the centered coefficient of determination (or centered R^2). When a specification contains a constant term,

$$\underbrace{\sum_{t=1}^{T}(y_t-\bar{y})^2}_{\text{Centered TSS}}=\underbrace{\sum_{t=1}^{T}(\hat{y}_t-\bar{\hat{y}})^2}_{\text{Centered RSS}}+\underbrace{\sum_{t=1}^{T}\hat{e}_t^2}_{\text{ESS}},$$

where $\bar{\hat{y}}=\bar{y}=\sum_{t=1}^{T}y_t/T$. Analogous to (3.6), the centered R^2 is defined as

$$\text{Centered } R^2=\frac{\text{Centered RSS}}{\text{Centered TSS}}=1-\frac{\text{ESS}}{\text{Centered TSS}}. \tag{3.7}$$

Centered R^2 also takes on values between 0 and 1 and is non-decreasing in the number of variables in the specification. In contrast with non-centered R^2, this measure *excludes* the effect of the constant term and hence is invariant with respect to constant addition.

When a specfication contains a constant term, we have

$$\sum_{t=1}^{T}(y_t-\bar{y})(\hat{y}_t-\bar{y})=\sum_{t=1}^{T}(\hat{y}_t-\bar{y}+\hat{e}_t)(\hat{y}_t-\bar{y})=\sum_{t=1}^{T}(\hat{y}_t-\bar{y})^2,$$

because $\sum_{t=1}^{T}\hat{y}_t\hat{e}_t=\sum_{t=1}^{t}\hat{e}_t=0$ by Theorem 3.2. It follows that

$$R^2=\frac{\sum_{t=1}^{T}(\hat{y}_t-\bar{y})^2}{\sum_{t=1}^{T}(y_t-\bar{y})^2}=\frac{[\sum_{t=1}^{T}(y_t-\bar{y})(\hat{y}_t-\bar{y})]^2}{[\sum_{t=1}^{T}(y_t-\bar{y})^2][\sum_{t=1}^{T}(\hat{y}_t-\bar{y})^2]}.$$

That is, the centered R^2 is also the squared sample correlation coefficient of y_t and $\hat{y}_t$, also known as the *squared multiple correlation coefficient*. If a specification does *not* contain a constant term, the centered R^2 may be negative; see Exercise 3.10.

Both centered and non-centered R^2 are still non-decreasing in the number of regressors. As such, if one try to determine a specification based on their R^2, the specification

with more regressors would be chosen. A modified measure is the adjusted R^2, $\bar{R}^2$, which is the centered R^2 adjusted for the degrees of freedom:

$$\bar{R}^2 = 1 - \frac{\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}/(T-k)}{(\boldsymbol{y}'\boldsymbol{y} - T\bar{y}^2)/(T-1)}.$$

This measure can also be expressed in different forms:

$$\bar{R}^2 = 1 - \frac{T-1}{T-k}(1-R^2) = R^2 - \frac{k-1}{T-k}(1-R^2).$$

That is, $\bar{R}^2$ is the centered R^2 with a penalty term depending on model complexity and explanatory ability. Observe that when k increases, $(k-1)/(T-k)$ increases but $1-R^2$ decreases. Whether the penalty term is larger or smaller depends on the trade-off between these two terms. Thus, $\bar{R}^2$ need not be increasing with the number of explanatory variables. Clearly, $\bar{R}^2 < R^2$ except for $k = 1$ or $R^2 = 1$. It can also be verified that $\bar{R}^2 < 0$ when $R^2 < (k-1)/(T-1)$.

Remark: As different dependent variables have different TSS, the associated specifications are therefore not comparable in terms of their R^2. For example, R^2 of the specifications with y and $\log y$ as dependent variables are not comparable.

3.3 Statistical Properties of the OLS Estimators

Readers should have noticed that the previous results, which are either algebraic or geometric, hold regardless of the random nature of data. To derive the statistical properties of the OLS estimator, some probabilistic conditions must be imposed.

3.3.1 Classical Conditions

The following conditions on data are usually known as the *classical conditions*.

[A1] $\boldsymbol{X}$ is non-stochastic.

[A2] $\boldsymbol{y}$ is a random vector such that

- **(i)** $\mathbb{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}_o$ for some $\boldsymbol{\beta}_o$;
- **(ii)** $\mathrm{var}(\boldsymbol{y}) = \sigma_o^2 \boldsymbol{I}_T$ for some $\sigma_o^2 > 0$.

[A3] $\boldsymbol{y}$ is a random vector such that $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}_o, \sigma_o^2 \boldsymbol{I}_T)$ for some $\boldsymbol{\beta}_o$ and $\sigma_o^2 > 0$.

Condition [A1] is not crucial, but it is quite convenient for subsequent analysis. Note that $\mathbb{E}(\boldsymbol{y})$ is the "averaging" behavior of y and may be interpreted as a systematic component of y. [A2](i) thus ensures that the postulated linear function $\boldsymbol{X}\boldsymbol{\beta}$ is a specification of this systematic component, correct up to unknown parameters. Condition [A2](ii) regulates that the variance-covariance matrix of $\boldsymbol{y}$ depends only on one parameter σ_o^2; such a matrix is also known as a *scalar covariance matrix*. Under [A2](ii), y_t, $t = 1, \ldots, T$, have the constant variance σ_o^2 and are pairwise uncorrelated (but not necessarily independent). Although conditions [A2] and [A3] impose the same structures for the mean and variance of $\boldsymbol{y}$, the latter is much stronger because it also specifies the distribution of $\boldsymbol{y}$. We have seen in Section 2.3 that uncorrelated normal random variables are also independent. Therefore, y_t, $t = 1, \ldots, T$, are i.i.d. (independently and identically distributed) normal random variables under [A3]. The linear specification (3.1) with [A1] and [A2] is known as the *classical linear model*, and (3.1) with [A1] and [A3] is also known as the *classical normal linear model*. The limitations of these conditions will be discussed in Section 3.7.

In addition to $\hat{\boldsymbol{\beta}}_T$, the new unknown parameter $\mathrm{var}(y_t) = \sigma_o^2$ in [A2](ii) and [A3] should be estimated as well. The OLS estimator for σ_o^2 is

$$\hat{\sigma}_T^2 = \frac{\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}}{T-k} = \frac{1}{T-k}\sum_{t=1}^{T}\hat{e}_t^2, \tag{3.8}$$

where k is the number of regressors. While $\hat{\boldsymbol{\beta}}_T$ is a *linear estimator* in the sense that it is a linear transformation of $\boldsymbol{y}$, $\hat{\sigma}_T^2$ is not. In the sections below we will derive the properties of the OLS estimators $\hat{\boldsymbol{\beta}}_T$ and $\hat{\sigma}_T^2$ under these classical conditions.

3.3.2 Without the Normality Condition

Under the imposed classical conditions, the OLS estimators have the following statistical properties.

Theorem 3.4 *Consider the linear specification* (3.1).

(a) *Given* [A1] *and* [A2](i), $\hat{\boldsymbol{\beta}}_T$ *is unbiased for* $\boldsymbol{\beta}_o$.

(b) *Given* [A1] *and* [A2], $\hat{\sigma}_T^2$ *is unbiased for* σ_o^2.

(c) *Given* [A1] *and* [A2], $\mathrm{var}(\hat{\boldsymbol{\beta}}_T) = \sigma_o^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

Proof: Given [A1] and [A2](i), $\hat{\boldsymbol{\beta}}_T$ is unbiased because

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_T) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathbb{E}(\boldsymbol{y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}_o = \boldsymbol{\beta}_o.$$

To prove (b), recall that $(\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{X} = \boldsymbol{0}$ so that the OLS residual vector can be written as

$$\hat{\boldsymbol{e}} = (\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{y} = (\boldsymbol{I}_T - \boldsymbol{P})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o).$$

Then, $\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)'(\boldsymbol{I}_T - \boldsymbol{P})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)$ which is a scalar, and

$$\begin{aligned}\mathbb{E}(\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}) &= \mathbb{E}[\operatorname{trace}\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)'(\boldsymbol{I}_T - \boldsymbol{P})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)\}] \\ &= \mathbb{E}[\operatorname{trace}\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)'(\boldsymbol{I}_T - \boldsymbol{P})\}].\end{aligned}$$

By interchanging the trace and expectation operators, we have from [A2](ii) that

$$\begin{aligned}\mathbb{E}(\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}) &= \operatorname{trace}\{\mathbb{E}[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)'(\boldsymbol{I}_T - \boldsymbol{P})]\} \\ &= \operatorname{trace}\{\mathbb{E}[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)'](\boldsymbol{I}_T - \boldsymbol{P})\} \\ &= \operatorname{trace}\{\sigma_o^2\boldsymbol{I}_T(\boldsymbol{I}_T - \boldsymbol{P})\} \\ &= \sigma_o^2 \operatorname{trace}(\boldsymbol{I}_T - \boldsymbol{P}).\end{aligned}$$

By Lemmas 1.12 and 1.14, $\operatorname{trace}(\boldsymbol{I}_T - \boldsymbol{P}) = \operatorname{rank}(\boldsymbol{I}_T - \boldsymbol{P}) = T - k$. Consequently,

$$\mathbb{E}(\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}) = \sigma_o^2(T - k),$$

so that

$$\mathbb{E}(\hat{\sigma}_T^2) = \mathbb{E}(\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}})/(T - k) = \sigma_o^2.$$

This proves the unbiasedness of $\hat{\sigma}_T^2$. Given that $\hat{\boldsymbol{\beta}}_T$ is a linear transformation of $\boldsymbol{y}$, we have from Lemma 2.4 that

$$\begin{aligned}\operatorname{var}(\hat{\boldsymbol{\beta}}_T) &= \operatorname{var}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}) \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma_o^2\boldsymbol{I}_T)\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma_o^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.\end{aligned}$$

This establishes (c). □

It can be seen that the unbiasedness of $\hat{\boldsymbol{\beta}}_T$ does not depend on [A2](ii), the variance property of $\boldsymbol{y}$. It is also clear that when $\hat{\sigma}_T^2$ is unbiased, the estimator

$$\widehat{\operatorname{var}}(\hat{\boldsymbol{\beta}}_T) = \hat{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

is also unbiased for $\operatorname{var}(\hat{\boldsymbol{\beta}}_T)$. The result below, known as the *Gauss-Markov theorem*, indicates that when [A1] and [A2] hold, $\hat{\boldsymbol{\beta}}_T$ is not only unbiased but also the best (most efficient) among all linear unbiased estimators for $\boldsymbol{\beta}_o$.

Theorem 3.5 (Gauss-Markov) *Given the linear specification* (3.1)*, suppose that* [A1] *and* [A2] *hold. Then the OLS estimator $\hat{\boldsymbol{\beta}}_T$ is the best linear unbiased estimator (BLUE) for $\boldsymbol{\beta}_o$.*

Proof: Consider an arbitrary linear estimator $\check{\boldsymbol{\beta}}_T = \boldsymbol{A}\boldsymbol{y}$, where $\boldsymbol{A}$ is non-stochastic. Writing $\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{C}$, $\check{\boldsymbol{\beta}}_T = \hat{\boldsymbol{\beta}}_T + \boldsymbol{C}\boldsymbol{y}$. Then,

$$\operatorname{var}(\check{\boldsymbol{\beta}}_T) = \operatorname{var}(\hat{\boldsymbol{\beta}}_T) + \operatorname{var}(\boldsymbol{C}\boldsymbol{y}) + 2\operatorname{cov}(\hat{\boldsymbol{\beta}}_T, \boldsymbol{C}\boldsymbol{y}).$$

By [A1] and [A2](i),

$$\mathbb{E}(\check{\boldsymbol{\beta}}_T) = \boldsymbol{\beta}_o + \boldsymbol{C}\boldsymbol{X}\boldsymbol{\beta}_o.$$

Since $\boldsymbol{\beta}_o$ is arbitrary, this estimator would be unbiased if, and only if, $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{0}$. This property further implies that

$$\begin{aligned}\operatorname{cov}(\hat{\boldsymbol{\beta}}_T, \boldsymbol{C}\boldsymbol{y}) &= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)\boldsymbol{y}'\boldsymbol{C}'] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathbb{E}[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)\boldsymbol{y}']\boldsymbol{C}' \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma_o^2\boldsymbol{I}_T)\boldsymbol{C}' \\ &= \boldsymbol{0}.\end{aligned}$$

Thus,

$$\operatorname{var}(\check{\boldsymbol{\beta}}_T) = \operatorname{var}(\hat{\boldsymbol{\beta}}_T) + \operatorname{var}(\boldsymbol{C}\boldsymbol{y}) = \operatorname{var}(\hat{\boldsymbol{\beta}}_T) + \sigma_o^2\boldsymbol{C}\boldsymbol{C}',$$

where $\sigma_o^2\boldsymbol{C}\boldsymbol{C}'$ is clearly a positive semi-definite matrix. This shows that for any linear unbiased estimator $\check{\boldsymbol{\beta}}_T$, $\operatorname{var}(\check{\boldsymbol{\beta}}_T) - \operatorname{var}(\hat{\boldsymbol{\beta}}_T)$ is positive semi-definite, so that $\hat{\boldsymbol{\beta}}_T$ is more efficient. □

Example 3.6 Given the data $[\boldsymbol{y}\ \boldsymbol{X}]$, where $\boldsymbol{X}$ is a nonstochastic matrix and can be partitioned as $[\boldsymbol{X}_1\ \boldsymbol{X}_2]$. Suppose that $\mathbb{E}(\boldsymbol{y}) = \boldsymbol{X}_1\boldsymbol{b}_1$ for some $\boldsymbol{b}_1$ and $\operatorname{var}(\boldsymbol{y}) = \sigma_o^2\boldsymbol{I}_T$ for some $\sigma_o^2 > 0$. Consider first the specification that contains only $\boldsymbol{X}_1$ but not $\boldsymbol{X}_2$:

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{e}.$$

Let $\hat{\boldsymbol{b}}_{1,T}$ denote the resulting OLS estimator. It is clear that $\hat{\boldsymbol{b}}_{1,T}$ is still a linear estimator and unbiased for $\boldsymbol{b}_1$ by Theorem 3.4(a). Moreover, it is the BLUE for $\boldsymbol{b}_1$ by Theorem 3.5 with the variance-covariance matrix

$$\operatorname{var}(\hat{\boldsymbol{b}}_{1,T}) = \sigma_o^2(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1},$$

by Theorem 3.4(c).

Consider now the linear specification that involves both $\boldsymbol{X}_1$ and irrelevant regressors $\boldsymbol{X}_2$:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}.$$

Thus, this specification cannot be a correct specification unless some of the parameters ($\boldsymbol{\beta}_2$) are restricted to zero. Let $\hat{\boldsymbol{\beta}}_T = (\hat{\boldsymbol{\beta}}_{1,T}' \; \hat{\boldsymbol{\beta}}_{2,T}')'$ be the OLS estimator of $\boldsymbol{\beta}$. Using Theorem 3.3, we find

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\beta}}_{1,T}) &= \mathbb{E}\left([\boldsymbol{X}_1'(\boldsymbol{I}_T - \boldsymbol{P}_2)\boldsymbol{X}_1]^{-1}\boldsymbol{X}_1'(\boldsymbol{I}_T - \boldsymbol{P}_2)\boldsymbol{y}\right) = \boldsymbol{b}_1, \\
\mathbb{E}(\hat{\boldsymbol{\beta}}_{2,T}) &= \mathbb{E}\left([\boldsymbol{X}_2'(\boldsymbol{I}_T - \boldsymbol{P}_1)\boldsymbol{X}_2]^{-1}\boldsymbol{X}_2'(\boldsymbol{I}_T - \boldsymbol{P}_1)\boldsymbol{y}\right) = \boldsymbol{0},
\end{aligned}$$

where $\boldsymbol{P}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$ and $\boldsymbol{P}_2 = \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'$. This shows that $\hat{\boldsymbol{\beta}}_T$ is unbiased for $(\boldsymbol{b}_1' \; \boldsymbol{0}')'$. Also,

$$\begin{aligned}
\operatorname{var}(\hat{\boldsymbol{\beta}}_{1,T}) &= \operatorname{var}([\boldsymbol{X}_1'(\boldsymbol{I}_T - \boldsymbol{P}_2)\boldsymbol{X}_1]^{-1}\boldsymbol{X}_1'(\boldsymbol{I}_T - \boldsymbol{P}_2)\boldsymbol{y}) \\
&= \sigma_o^2[\boldsymbol{X}_1'(\boldsymbol{I}_T - \boldsymbol{P}_2)\boldsymbol{X}_1]^{-1}.
\end{aligned}$$

Given that $\boldsymbol{P}_2$ is a positive semi-definite matrix,

$$\boldsymbol{X}_1'\boldsymbol{X}_1 - \boldsymbol{X}_1'(\boldsymbol{I}_T - \boldsymbol{P}_2)\boldsymbol{X}_1 = \boldsymbol{X}_1'\boldsymbol{P}_2\boldsymbol{X}_1,$$

must also be positive semi-definite. It follows from Lemma 1.9 that

$$[\boldsymbol{X}_1'(\boldsymbol{I}_T - \boldsymbol{P}_2)\boldsymbol{X}_1]^{-1} - (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}$$

is a positive semi-definite matrix. This shows that $\hat{\boldsymbol{b}}_{1,T}$ is more efficient than $\hat{\boldsymbol{\beta}}_{1,T}$, as it ought to be. When $\boldsymbol{X}_1'\boldsymbol{X}_2 = \boldsymbol{0}$, i.e., the columns of $\boldsymbol{X}_1$ are orthogonal to the columns of $\boldsymbol{X}_2$, we immediately have $(\boldsymbol{I}_T - \boldsymbol{P}_2)\boldsymbol{X}_1 = \boldsymbol{X}_1$, so that $\hat{\boldsymbol{\beta}}_{1,T} = \hat{\boldsymbol{b}}_{1,T}$. In this case, estimating a more complex specification does not result in efficiency loss. □

Remark: The Gauss-Markov theorem does not apply to the estimators for the specification $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}$ because, unlike [A2](i), the true parameter vector $\boldsymbol{\beta}_o = (\boldsymbol{b}_1' \; \boldsymbol{0}')'$ is not arbitrary but involves the restriction that some of its elements must be zero. This example thus shows that when this restriction is not taken into account, the resulting OLS estimator, while being unbiased, is no longer the most efficient.

3.3.3 With the Normality Condition

We have learned that the normality condition [A3] is much stronger than [A2]. With this stronger condition, more can be said about the OLS estimators.

Theorem 3.7 *Given the linear specification* (3.1)*, suppose that* [A1] *and* [A3] *hold.*

(a) $\hat{\boldsymbol{\beta}}_T \sim N(\boldsymbol{\beta}_o, \sigma_o^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$.

(b) $(T-k)\hat{\sigma}_T^2/\sigma_o^2 \sim \chi^2(T-k)$.

(c) $\hat{\sigma}_T^2$ *has mean* σ_o^2 *and variance* $2\sigma_o^4/(T-k)$.

Proof: As $\hat{\boldsymbol{\beta}}_T$ is a linear transformation of $\boldsymbol{y}$, it is also normally distributed as

$$\hat{\boldsymbol{\beta}}_T \sim N(\boldsymbol{\beta}_o, \sigma_o^2(\boldsymbol{X}'\boldsymbol{X})^{-1}),$$

by Lemma 2.6, where its mean and variance-covariance matrix are as in Theorem 3.4(a) and (c). To prove the assertion (b), we again write $\hat{\boldsymbol{e}} = (\boldsymbol{I}_T - \boldsymbol{P})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)$ and deduce

$$(T-k)\hat{\sigma}_T^2/\sigma_o^2 = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}/\sigma_o^2 = \boldsymbol{y}^{*\prime}(\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{y}^*,$$

where $\boldsymbol{y}^* = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)/\sigma_o$. Let $\boldsymbol{C}$ be the orthogonal matrix that diagonalizes the symmetric and idempotent matrix $\boldsymbol{I}_T - \boldsymbol{P}$. Then, $\boldsymbol{C}'(\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{C} = \boldsymbol{\Lambda}$. Since $\mathrm{rank}(\boldsymbol{I}_T - \boldsymbol{P}) = T-k$, $\boldsymbol{\Lambda}$ contains $T-k$ eigenvalues equal to one and k eigenvalues equal to zero by Lemma 1.11. Without loss of generality we can write

$$\boldsymbol{y}^{*\prime}(\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{y}^* = \boldsymbol{y}^{*\prime}\boldsymbol{C}[\boldsymbol{C}'(\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{C}]\boldsymbol{C}'\boldsymbol{y}^* = \boldsymbol{\eta}' \begin{bmatrix} \boldsymbol{I}_{T-k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{\eta},$$

where $\boldsymbol{\eta} = \boldsymbol{C}'\boldsymbol{y}^*$. Again by Lemma 2.6, $\boldsymbol{y}^* \sim N(\boldsymbol{0}, \boldsymbol{I}_T)$ under [A3]. Hence, $\boldsymbol{\eta} \sim N(\boldsymbol{0}, \boldsymbol{I}_T)$, so that η_i are independent, standard normal random variables. Consequently,

$$\boldsymbol{y}^{*\prime}(\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{y}^* = \sum_{i=1}^{T-k} \eta_i^2 \sim \chi^2(T-k).$$

This proves (b). Noting that the mean of $\chi^2(T-k)$ is $T-k$ and variance is $2(T-k)$, the assertion (c) is just a direct consequence of (b). □

Suppose that we believe that [A3] is true and specify the log-likelihood function of $\boldsymbol{y}$ as:

$$\log L(\boldsymbol{\beta}, \sigma^2) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

The first order conditions of maximizing this log-likelihood are

$$\nabla_{\boldsymbol{\beta}} \log L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \mathbf{0},$$
$$\nabla_{\sigma^2} \log L(\boldsymbol{\beta}, \sigma^2) = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = 0,$$

and their solutions are the MLEs $\tilde{\boldsymbol{\beta}}_T$ and $\tilde{\sigma}_T^2$. The first k equations above are equivalent to the OLS normal equations (3.3). It follows that the OLS estimator $\hat{\boldsymbol{\beta}}_T$ is also the MLE $\tilde{\boldsymbol{\beta}}_T$. Plugging $\hat{\boldsymbol{\beta}}_T$ into the first order conditions we can solve for σ^2 and obtain

$$\tilde{\sigma}_T^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_T)'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_T)}{T} = \frac{\hat{e}'\hat{e}}{T}, \tag{3.9}$$

which is different from the OLS variance estimator (3.8).

The conclusion below is stronger than the Gauss-Markov theorem (Theorem 3.5).

Theorem 3.8 *Given the linear specification* (3.1)*, suppose that* [A1] *and* [A3] *hold. Then the OLS estimators* $\hat{\boldsymbol{\beta}}_T$ *and* $\hat{\sigma}_T^2$ *are the best unbiased estimators for* $\boldsymbol{\beta}_o$ *and* σ_o^2*, respectively.*

Proof: The score vector is

$$\boldsymbol{s}(\boldsymbol{\beta}, \sigma^2) = \begin{bmatrix} \frac{1}{\sigma^2}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \end{bmatrix},$$

and the Hessian matrix of the log-likelihood function is

$$\begin{bmatrix} -\frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X} & -\frac{1}{\sigma^4}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ -\frac{1}{\sigma^4}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{X} & \frac{T}{2\sigma^4} - \frac{1}{\sigma^6}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \end{bmatrix}.$$

It is easily verified that when [A3] is true, $\mathbb{E}[\boldsymbol{s}(\boldsymbol{\beta}_o, \sigma_o^2)] = \mathbf{0}$, and the expected value of the Hessian matrix evaluated at $\boldsymbol{\beta}_o$ and σ_o^2 is

$$\begin{bmatrix} -\frac{1}{\sigma_o^2}\boldsymbol{X}'\boldsymbol{X} & \mathbf{0} \\ \mathbf{0} & -\frac{T}{2\sigma_o^4} \end{bmatrix}.$$

The information matrix equality (Lemma 2.9) ensures that the negative of this matrix equals the information matrix. The inverse of the information matrix is then

$$\begin{bmatrix} \sigma_o^2(\boldsymbol{X}'\boldsymbol{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2\sigma_o^4}{T} \end{bmatrix},$$

which is the Cramér-Rao lower bound by Lemma 2.10. Clearly, $\text{var}(\hat{\boldsymbol{\beta}}_T)$ achieves this lower bound so that $\hat{\boldsymbol{\beta}}_T$ must be the best unbiased estimator for $\boldsymbol{\beta}_o$. Although the variance of $\hat{\sigma}_T^2$ is greater than the lower bound, it can be shown that $\hat{\sigma}_T^2$ is still the best unbiased estimator for σ_o^2; see Rao (1973,) for a proof. □

Remark: Comparing to the Gauss-Markov theorem, Theorem 3.8 gives a stronger result at the expense of a stronger condition (the normality condition [A3]). The OLS estimators now are the best (most efficient) in a much larger class of estimators, namely, the class of unbiased estimators. Note also that Theorem 3.8 covers $\hat{\sigma}_T^2$, whereas the Gauss-Markov theorem does not.

3.4 Hypotheses Testing

After a specification is estimated, it is often desirable to test various economic and econometric hypotheses. Given the classical conditions [A1] and [A3], we consider the linear hypothesis

$$\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r}, \tag{3.10}$$

where $\boldsymbol{R}$ is a $q \times k$ non-stochastic matrix with rank $q < k$, and $\boldsymbol{r}$ is a vector of pre-specified, hypothetical values.

3.4.1 Tests for Linear Hypotheses

If the null hypothesis (3.10) is true, it is reasonable to expect that $\boldsymbol{R}\hat{\boldsymbol{\beta}}_T$ is "close" to the hypothetical value $\boldsymbol{r}$; otherwise, they should be quite different. Here, the closeness between $\boldsymbol{R}\hat{\boldsymbol{\beta}}_T$ and $\boldsymbol{r}$ must be justified by the null distribution of the test statistics.

If there is only a single hypothesis, the null hypothesis (3.10) is such that $\boldsymbol{R}$ is a row vector ($q = 1$) and $\boldsymbol{r}$ is a scalar. Note that a single hypothesis may involve two or more parameters. Consider the following statistic:

$$\frac{\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}}{\sigma_o[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{1/2}}.$$

By Theorem 3.7(a), $\hat{\boldsymbol{\beta}}_T \sim N(\boldsymbol{\beta}_o, \sigma_o^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$, and hence

$$\boldsymbol{R}\hat{\boldsymbol{\beta}}_T \sim N(\boldsymbol{R}\boldsymbol{\beta}_o, \sigma_o^2\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}').$$

Under the null hypothesis, we have

$$\frac{\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}}{\sigma_o[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{1/2}} = \frac{\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o)}{\sigma_o[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{1/2}} \sim N(0,1). \tag{3.11}$$

Although the left-hand side has a known distribution, it cannot be used as a test statistic because σ_o is unknown. Replacing σ_o by its OLS estimator $\hat{\sigma}_T$ yields an operational statistic:

$$\tau = \frac{\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}}{\hat{\sigma}_T[R(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{1/2}}. \tag{3.12}$$

The null distribution of τ is given in the result below.

Theorem 3.9 *Given the linear specification* (3.1)*, suppose that* [A1] *and* [A3] *hold. Then under the null hypothesis* (3.10) *with* $\boldsymbol{R}$ *a* $1 \times k$ *vector,*

$$\tau \sim t(T-k),$$

where τ *is given by* (3.12)*.*

Proof: We first write the statistic τ as

$$\tau = \frac{\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}}{\sigma_o[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{1/2}} \Bigg/ \sqrt{\frac{(T-k)\hat{\sigma}_T^2/\sigma_o^2}{T-k}},$$

where the numerator is distributed as $N(0,1)$ by (3.11), and $(T-k)\hat{\sigma}_T^2/\sigma_o^2$ is distributed as $\chi^2(T-k)$ by Theorem 3.7(b). Hence, the square of the denominator is a central χ^2 random variable divided by its degrees of freedom $T-k$. The assertion follows if we can show that the numerator and denominator are independent. Note that the random components of the numerator and denominator are, respectively, $\hat{\boldsymbol{\beta}}_T$ and $\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$, where $\hat{\boldsymbol{\beta}}_T$ and $\hat{\boldsymbol{e}}$ are two normally distributed random vectors with the covariance matrix

$$\begin{aligned}
\operatorname{cov}(\hat{\boldsymbol{e}}, \hat{\boldsymbol{\beta}}_T) &= \mathbb{E}[(\boldsymbol{I}_T - \boldsymbol{P})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)\boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}] \\
&= (\boldsymbol{I}_T - \boldsymbol{P})\,\mathbb{E}[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)\boldsymbol{y}']\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma_o^2(\boldsymbol{I}_T - \boldsymbol{P})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \boldsymbol{0}.
\end{aligned}$$

Since uncorrelated normal random vectors are also independent, $\hat{\boldsymbol{\beta}}_T$ is independent of $\hat{\boldsymbol{e}}$. By Lemma 2.1, we conclude that $\hat{\boldsymbol{\beta}}_T$ is also independent of $\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$. □

The statistic τ is known as the t statistic, and its null distribution is $t(T-k)$ by Theorem 3.9. When the alternative hypothesis is $\boldsymbol{R}\boldsymbol{\beta}_o \neq \boldsymbol{r}$, this is a two-sided test; when the alternative hypothesis is $\boldsymbol{R}\boldsymbol{\beta}_o > \boldsymbol{r}$ (or $\boldsymbol{R}\boldsymbol{\beta}_o < \boldsymbol{r}$), this is a one-sided test. For each test, we first choose a small significance level α and then determine the critical

region C_α. For the two-sided t test, we can find the values $\pm t_{\alpha/2}(T-k)$ from the t table such that

$$\begin{aligned}\alpha &= \mathbb{P}\{\tau < -t_{\alpha/2}(T-k) \text{ or } \tau > t_{\alpha/2}(T-k)\} \\ &= 1 - \mathbb{P}\{-t_{\alpha/2}(T-k) \le \tau \le t_{\alpha/2}(T-k)\}.\end{aligned}$$

The critical region is then

$$C_\alpha = (-\infty, -t_{\alpha/2}(T-k)) \cup (t_{\alpha/2}(T-k), \infty),$$

and $\pm t_{\alpha/2}(T-k)$ are the critical values at the significance level α. For the alternative hypothesis $\boldsymbol{R}\boldsymbol{\beta}_o > \boldsymbol{r}$, the critical region is $(t_\alpha(T-k), \infty)$, where $t_\alpha(T-k)$ is the critical value such that

$$\alpha = \mathbb{P}\{\tau > t_\alpha(T-k)\}.$$

Similarly, for the alternative $\boldsymbol{R}\boldsymbol{\beta}_o < \boldsymbol{r}$, the critical region is $(-\infty, -t_\alpha(T-k))$.

The null hypothesis is rejected at the significance level α when τ falls in the critical region. As α is small, the event $\{\tau \in C_\alpha\}$ is unlikely under the null hypothesis. When τ does take an extreme value relative to the critical values, it is an evidence against the null hypothesis. The decision of rejecting the null hypothesis could be wrong, but the probability of the type I error will not exceed α. When τ takes a "reasonable" value in the sense that it falls in the complement of the critical region, the null hypothesis is not rejected.

Example 3.10 To test a single coefficient equal to zero: $\beta_i = 0$, we choose $\boldsymbol{R}$ as the transpose of the ith Cartesian unit vector:

$$\boldsymbol{R} = [\,0 \quad \cdots \quad 0 \quad 1 \quad 0 \quad \cdots \quad 0\,].$$

Let m^{ii} be the i^{th} diagonal element of $\boldsymbol{M}^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$. Then, $\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}' = m^{ii}$. The t statistic for this hypothesis, also known as the t ratio, is

$$\tau = \frac{\hat{\beta}_{i,T}}{\hat{\sigma}_T\sqrt{m^{ii}}} \sim t(T-k).$$

When a t ratio rejects the null hypothesis, it is said that the corresponding estimated coefficient is significantly different from zero; econometric packages usually report t ratios along with the coefficient estimates. □

Example 3.11 To test the single hypothesis $\beta_i + \beta_j = 0$, we set $\boldsymbol{R}$ as

$$\boldsymbol{R} = [\, 0 \quad \cdots \quad 0 \quad 1 \quad 0 \quad \cdots \quad 0 \quad 1 \quad 0 \quad \cdots \quad 0 \,].$$

Hence, $\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}' = m^{ii} + 2m^{ij} + m^{jj}$, where m^{ij} is the $(i,j)^{\text{th}}$ element of $\boldsymbol{M}^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$. The t statistic is

$$\tau = \frac{\hat{\beta}_{i,T} + \hat{\beta}_{j,T}}{\hat{\sigma}_T (m^{ii} + 2m^{ij} + m^{jj})^{1/2}} \sim t(T-k). \quad \square$$

Several hypotheses can also be tested jointly. Consider the null hypothesis $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r}$, where $\boldsymbol{R}$ is now a $q \times k$ matrix ($q \geq 2$) and $\boldsymbol{r}$ is a vector. This hypothesis involves q single hypotheses. Similar to (3.11), we have under the null hypothesis that

$$[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})/\sigma_o \sim N(\boldsymbol{0}, \boldsymbol{I}_q).$$

Therefore,

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})/\sigma_o^2 \sim \chi^2(q). \tag{3.13}$$

Again, we can replace σ_o^2 by its OLS estimator $\hat{\sigma}_T^2$ to obtain an operational statistic:

$$\varphi = \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})}{\hat{\sigma}_T^2 q}. \tag{3.14}$$

The next result gives the null distribution of φ.

Theorem 3.12 *Given the linear specification* (3.1)*, suppose that* [A1] *and* [A3] *hold. Then under the null hypothesis* (3.10) *with* $\boldsymbol{R}$ *a* $q \times k$ *matrix with rank* $q < k$*, we have*

$$\varphi \sim F(q, T-k),$$

where φ *is given by* (3.14)*.*

Proof: Note that

$$\varphi = \frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})/(\sigma_o^2 q)}{(T-k)\frac{\hat{\sigma}_T^2}{\sigma_o^2}\Big/(T-k)}.$$

In view of (3.13) and the proof of Theorem 3.9, the numerator and denominator terms are two independent χ^2 random variables, each divided by its degrees of freedom. The assertion follows from the definition of F random variable. $\square$

The statistic φ is known as the F statistic. We reject the null hypothesis at the significance level α when φ is too large relative to the critical value $F_\alpha(q, T-k)$ of the F table, where $F_\alpha(q, T-k)$ is such that

$$\alpha = \mathbb{P}\{\varphi > F_\alpha(q, T-k)\}.$$

If there is only a single hypothesis, the F statistic is just the square of the corresponding t statistic. When φ rejects the null hypothesis, it simply suggests that there is evidence against at least one single hypothesis. The inference of a joint test is, however, not necessary the same as the inference of individual tests; see also Section 3.5.

Example 3.13 Joint null hypothesis: H_o: $\beta_1 = b_1$ and $\beta_2 = b_2$. The F statistic is

$$\varphi = \frac{1}{2\hat{\sigma}_T^2} \begin{pmatrix} \hat{\beta}_{1,T} - b_1 \\ \hat{\beta}_{2,T} - b_2 \end{pmatrix}' \begin{bmatrix} m^{11} & m^{12} \\ m^{21} & m^{22} \end{bmatrix}^{-1} \begin{pmatrix} \hat{\beta}_{1,T} - b_1 \\ \hat{\beta}_{2,T} - b_2 \end{pmatrix} \sim F(2, T-k),$$

where m^{ij} is as defined in Example 3.11. □

Remark: For the null hypothesis of s coefficients being zero, if the corresponding F statistic $\varphi > 1$ ($\varphi < 1$), dropping these s regressors will reduce (increase) $\bar{R}^2$; see Exercise 3.12.

3.4.2 Power of the Tests

Recall that the power of a test is the probability of rejecting the null hypothesis when the null hypothesis is indeed false. In this section, we consider the hypothesis $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r} + \boldsymbol{\delta}$, where $\boldsymbol{\delta}$ characterizes the deviation from the null hypothesis, and analyze the power performance of the t and F tests.

Theorem 3.14 *Given the linear specification* (3.1)*, suppose that* [A1] *and* [A3] *hold. Then under the hypothesis that* $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r} + \boldsymbol{\delta}$*, where* $\boldsymbol{R}$ *is a* $q \times k$ *matrix with rank* $q < k$*, we have*

$$\varphi \sim F(q, T-k;\, \boldsymbol{\delta}'\boldsymbol{D}^{-1}\boldsymbol{\delta}, 0),$$

where φ *is given by* (3.14)*,* $\boldsymbol{D} = \sigma_o^2[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']$*, and* $\boldsymbol{\delta}'\boldsymbol{D}^{-1}\boldsymbol{\delta}$ *is the non-centrality parameter of the numerator term.*

Proof: When $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r} + \boldsymbol{\delta}$,

$$\begin{aligned} &[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})/\sigma_o \\ &\quad = [\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1/2}[\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o) + \boldsymbol{\delta}]/\sigma_o. \end{aligned}$$

Given [A3],

$$[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1/2}\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o)/\sigma_o. \sim N(\boldsymbol{0}, \boldsymbol{I}_q),$$

and hence

$$[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})/\sigma_o \sim N(\boldsymbol{D}^{-1/2}\boldsymbol{\delta}, \boldsymbol{I}_q).$$

It follows from Lemma 2.7 that

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})/\sigma_o^2 \sim \chi^2(q;\, \boldsymbol{\delta}'\boldsymbol{D}^{-1}\boldsymbol{\delta}),$$

which is the non-central χ^2 distribution with q degrees of freedom and the non-centrality parameter $\boldsymbol{\delta}'\boldsymbol{D}^{-1}\boldsymbol{\delta}$. This is in contrast with (3.13) which has a central χ^2 distribution under the null hypothesis. As $(T-k)\hat{\sigma}_T^2/\sigma_o^2$ is still distributed as $\chi^2(T-k)$ by Theorem 3.7(b), the assertion follows because the numerator and denominator of φ are independent. □

Clearly, when the null hypothesis is correct, we have $\boldsymbol{\delta} = \boldsymbol{0}$, so that $\varphi \sim F(q, T-k)$. Theorem 3.14 thus includes Theorem 3.12 as a special case. In particular, for testing a single hypothesis, we have

$$\tau \sim t(T-k;\, \boldsymbol{D}^{-1/2}\boldsymbol{\delta}),$$

which reduces to $t(T-k)$ when $\boldsymbol{\delta} = 0$, as in Theorem 3.9.

Theorem 3.14 implies that when $\boldsymbol{R}\boldsymbol{\beta}_o$ deviates farther from the hypothetical value $\boldsymbol{r}$, the non-centrality parameter $\boldsymbol{\delta}'\boldsymbol{D}^{-1}\boldsymbol{\delta}$ increases, and so does the power. We illustrate this point using the following two examples, where the power are computed using the GAUSS program. For the null distribution $F(2, 20)$, the critical value at 5% level is 3.49. Then for $F(2, 20; \nu_1, 0)$ with the non-centrality parameter $\nu_1 = 1, 3, 5$, the probabilities that φ exceeds 3.49 are approximately 12.1%, 28.2%, and 44.3%, respectively. For the null distribution $F(5, 60)$, the critical value at 5% level is 2.37. Then for $F(5, 60; \nu_1, 0)$ with $\nu_1 = 1, 3, 5$, the probabilities that φ exceeds 2.37 are approximately 9.4%, 20.5%, and 33.2%, respectively. In both cases, the power increases with the non-centrality parameter.

3.4.3 An Alternative Approach

Given the specification (3.1), we may take the constraint $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r}$ into account and consider the *constrained* OLS estimation that finds the saddle point of the Lagrangian:

$$\min_{\boldsymbol{\beta},\boldsymbol{\lambda}} \frac{1}{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{r})'\boldsymbol{\lambda},$$

where $\boldsymbol{\lambda}$ is the $q \times 1$ vector of Lagrangian multipliers. It is straightforward to show that the solutions are

$$\begin{aligned} \ddot{\boldsymbol{\lambda}}_T &= 2[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}), \\ \ddot{\boldsymbol{\beta}}_T &= \hat{\boldsymbol{\beta}}_T - (\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}'\ddot{\boldsymbol{\lambda}}_T/2, \end{aligned} \tag{3.15}$$

which will be referred to as the constrained OLS estimators.

Given $\ddot{\boldsymbol{\beta}}_T$, the vector of constrained OLS residuals is

$$\ddot{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\ddot{\boldsymbol{\beta}}_T = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_T + \boldsymbol{X}(\hat{\boldsymbol{\beta}}_T - \ddot{\boldsymbol{\beta}}_T) = \hat{\boldsymbol{e}} + \boldsymbol{X}(\hat{\boldsymbol{\beta}}_T - \ddot{\boldsymbol{\beta}}_T).$$

It follows from (3.15) that

$$\begin{aligned} \hat{\boldsymbol{\beta}}_T - \ddot{\boldsymbol{\beta}}_T &= (\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}'\ddot{\boldsymbol{\lambda}}_T/2 \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}). \end{aligned}$$

The inner product of $\ddot{\boldsymbol{e}}$ is then

$$\begin{aligned} \ddot{\boldsymbol{e}}'\ddot{\boldsymbol{e}} &= \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} + (\hat{\boldsymbol{\beta}}_T - \ddot{\boldsymbol{\beta}}_T)'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}_T - \ddot{\boldsymbol{\beta}}_T) \\ &= \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} + (\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}). \end{aligned}$$

Note that the second term on the right-hand side is nothing but the numerator of the F statistic (3.14). The F statistic now can be written as

$$\varphi = \frac{\ddot{\boldsymbol{e}}'\ddot{\boldsymbol{e}} - \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}}{q\hat{\sigma}_T^2} = \frac{(\mathrm{ESS_c} - \mathrm{ESS_u})/q}{\mathrm{ESS_u}/(T-k)}, \tag{3.16}$$

where $\mathrm{ESS_c} = \ddot{\boldsymbol{e}}'\ddot{\boldsymbol{e}}$ and $\mathrm{ESS_u} = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$ denote, respectively, the ESS resulted from constrained and unconstrained estimations. Dividing the numerator and denominator of (3.16) by centered TSS ($\boldsymbol{y}'\boldsymbol{y} - T\bar{y}^2$) yields another equivalent expression for φ:

$$\varphi = \frac{(R_{\mathrm{u}}^2 - R_{\mathrm{c}}^2)/q}{(1 - R_{\mathrm{u}}^2)/(T-k)}, \tag{3.17}$$

where R_c^2 and R_u^2 are, respectively, the centered coefficient of determination of constrained and unconstrained estimations. As the numerator of (3.17), $R_{\mathrm{u}}^2 - R_{\mathrm{c}}^2$, can be interpreted as the loss of fit due to the imposed constraint, the F test is in effect a loss-of-fit test. The null hypothesis is rejected when the constrained specification fits data much worse.

Example 3.15 Consider the specification: $y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + e_t$. Given the hypothesis (constraint) $\beta_2 = \beta_3$, the resulting constrained specification is

$$y_t = \beta_1 + \beta_2(x_{t2} + x_{t3}) + e_t.$$

By estimating these two specifications separately, we obtain ESS_u and ESS_c, from which the F statistic can be easily computed. □

Example 3.16 Test the null hypothesis that all the coefficients (except the constant term) equal zero. The resulting constrained specification is $y_t = \beta_1 + e_t$, so that $R_\text{c}^2 = 0$. Then, (3.17) becomes

$$\varphi = \frac{R_\text{u}^2/(k-1)}{(1-R_\text{u}^2)/(T-k)} \sim F(k-1, T-k),$$

which requires only estimation of the unconstrained specification. This test statistic is also routinely reported by most of econometric packages and known as the "regression F test." □

3.5 Confidence Regions

In addition to point estimators for parameters, we may also be interested in finding confidence intervals for parameters. A confidence interval for $\beta_{i,o}$ with the confidence coefficient $(1-\alpha)$ is the interval $(\underline{g}_\alpha, \overline{g}_\alpha)$ that satisfies

$$\mathbb{P}\{\, \underline{g}_\alpha \le \beta_{i,o} \le \overline{g}_\alpha\} = 1 - \alpha.$$

That is, we are $(1-\alpha)\times 100$ percent sure that such an interval would include the true parameter $\beta_{i,o}$.

From Theorem 3.9, we know

$$\mathbb{P}\left\{ -t_{\alpha/2}(T-k) \le \frac{\hat\beta_{i,T} - \beta_{i,o}}{\hat\sigma_T \sqrt{m^{ii}}} \le t_{\alpha/2}(T-k) \right\} = 1-\alpha,$$

where m^{ii} is the $i^{\,\text{th}}$ diagonal element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, and $t_{\alpha/2}(T-k)$ is the critical value of the (two-sided) t test at the significance level α. Equivalently, we have

$$\mathbb{P}\left\{ \hat\beta_{i,T} - t_{\alpha/2}(T-k)\hat\sigma_T\sqrt{m^{ii}} \le \beta_{i,o} \le \hat\beta_{i,T} + t_{\alpha/2}(T-k)\hat\sigma_T\sqrt{m^{ii}} \right\} = 1-\alpha.$$

This shows that the confidence interval for $\beta_{i,o}$ can be constructed by setting

$$\underline{g}_\alpha = \hat\beta_{i,T} - t_{\alpha/2}(T-k)\hat\sigma_T\sqrt{m^{ii}},$$
$$\overline{g}_\alpha = \hat\beta_{i,T} + t_{\alpha/2}(T-k)\hat\sigma_T\sqrt{m^{ii}}.$$

It should be clear that the greater the confidence coefficient (i.e., α smaller), the larger is the magnitude of the critical values $\pm t_{\alpha/2}(T-k)$ and hence the resulting confidence interval.

The confidence region for $\boldsymbol{R}\boldsymbol{\beta}_o$ with the confidence coefficient $(1-\alpha)$ satisfies

$$\begin{aligned}&\mathbb{P}\{(\hat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}_o)'\boldsymbol{R}'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}_o)/(q\hat{\sigma}_T^2)\le F_\alpha(q,T-k)\}\\&\quad=1-\alpha,\end{aligned}$$

where $F_\alpha(q,T-k)$ is the critical value of the F test at the significance level α.

Example 3.17 The confidence region for $(\beta_{1,o}=b_1,\beta_{2,o}=b_2)$. Suppose $T-k=30$ and $\alpha=0.05$, then $F_{0.05}(2,30)=3.32$. In view of Example 3.13,

$$\mathbb{P}\left\{\frac{1}{2\hat{\sigma}_T^2}\begin{pmatrix}\hat{\beta}_{1,T}-b_1\\ \hat{\beta}_{2,T}-b_2\end{pmatrix}'\begin{bmatrix}m^{11}&m^{12}\\ m^{21}&m^{22}\end{bmatrix}^{-1}\begin{pmatrix}\hat{\beta}_{1,T}-b_1\\ \hat{\beta}_{2,T}-b_2\end{pmatrix}\le 3.32\right\}=0.95,$$

which results in an ellipse with the center $(\hat{\beta}_{1,T},\hat{\beta}_{2,T})$. □

Remark: A point $(\beta_{1,o},\beta_{2,o})$ may be outside the joint confidence ellipse but inside the confidence box formed by individual confidence intervals. Hence, each t ratio may show that the corresponding coefficient is insignificantly different from zero, while the F test indicates that both coefficients are not jointly insignificant. It is also possible that (β_1,β_2) is outside the confidence box but inside the joint confidence ellipse. That is, each t ratio may show that the corresponding coefficient is significantly different from zero, while the F test indicates that both coefficients are jointly insignificant. See also an illustrative example in Goldberger (1991, Chap. 19).

3.6 Multicollinearity

In Section 3.2.2 we have seen that a linear specification suffers from the problem of exact multicollinearity if the basic identifiability requirement (i.e., $\boldsymbol{X}$ is of full column rank) is not satisfied. In this case, the OLS estimator cannot be computed as (3.4). This problem may be avoided by modifying the postulated specifications.

3.6.1 Near Multicollinearity

In practice, it is more common that explanatory variables are related to some extent but do not satisfy an exact linear relationship. This is usually referred to as the problem of *near multicollinearity*. But as long as there is no exact multicollinearity, parameters can still be estimated by the OLS method, and the resulting estimator remains the BLUE under [A1] and [A2].

Nevertheless, there are still complaints about near multicollinearity in empirical studies. In some applications, parameter estimates are very sensitive to small changes in data. It is also possible that individual t ratios are all insignificant, but the regression F statistic is highly significant. These symptoms are usually attributed to near multicollinearity. This is not entirely correct, however. Write $\boldsymbol{X} = [\boldsymbol{x}_i \ \boldsymbol{X}_i]$, where $\boldsymbol{X}_i$ is the submatrix of $\boldsymbol{X}$ excluding the i^{th} column $\boldsymbol{x}_i$. By the result of Theorem 3.3, the variance of $\hat{\beta}_{i,T}$ can be expressed as

$$\text{var}(\hat{\beta}_{i,T}) = \text{var}([\boldsymbol{x}_i'(\boldsymbol{I} - \boldsymbol{P}_i)\boldsymbol{x}_i]^{-1}x_i'(\boldsymbol{I} - \boldsymbol{P}_i)\boldsymbol{y}) = \sigma_o^2[\boldsymbol{x}_i'(\boldsymbol{I} - \boldsymbol{P}_i)\boldsymbol{x}_i]^{-1},$$

where $\boldsymbol{P}_i = \boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'$. It can also be verified that

$$\text{var}(\hat{\beta}_{i,T}) = \frac{\sigma_o^2}{\sum_{t=1}^T (x_{ti} - \bar{x}_i)^2(1 - R^2(i))},$$

where $R^2(i)$ is the centered coefficient of determination from the auxiliary regression of $\boldsymbol{x}_i$ on $\boldsymbol{X}_i$. When $\boldsymbol{x}_i$ is closely related to other explanatory variables, $R^2(i)$ is high so that $\text{var}(\hat{\beta}_{i,T})$ would be large. This explains why $\hat{\beta}_{i,T}$ are sensitive to data changes and why corresponding t ratios are likely to be insignificant. Near multicollinearity is not a necessary condition for these problems, however. Large $\text{var}(\hat{\beta}_{i,T})$ may also arise due to small variations of x_{ti} and/or large σ_o^2.

Even when a large value of $\text{var}(\hat{\beta}_{i,T})$ is indeed resulted from high $R^2(i)$, there is nothing wrong statistically. It is often claimed that "severe multicollinearity can make an important variable look insignificant." As Goldberger (1991) correctly pointed out, this statement simply confuses statistical significance with economic importance. These large variances merely reflect the fact that parameters cannot be precisely estimated from the given data set.

Near multicollinearity is in fact a problem related to data and model specification. If it does cause problems in estimation and hypothesis testing, one may try to break the approximate linear relationship by, e.g., adding more observations to the data set (if plausible) or dropping some variables from the current specification. More sophisticated statistical methods, such as the ridge estimator and principal component regressions, may also be used; details of these methods can be found in other econometrics textbooks.

3.6.2 Digress: Dummy Variables

A linear specification may include some qualitative variables to indicate the presence or absence of certain attributes of the dependent variable. These qualitative variables are typically represented by *dummy variables* which classify data into different categories.

For example, let y_i denote the annual salary of college teacher i and x_i the years of teaching experience. Consider the dummy variable: $D_i = 1$ if i is a male and $D_i = 0$ if i is a female. Then, the specification

$$y_i = \alpha_0 + \alpha_1 D_i + \beta x_i + e_i$$

yields two regression lines with different intercepts. The "male" regression line has the intercept $\alpha_0 + \alpha_1$, and the "female" regression line has the intercept α_0. We may test the hypothesis $\alpha_1 = 0$ to see if there is a difference between the starting salaries of male and female teachers.

This specification can be expanded to incorporate an interaction term between D and x:

$$y_i = \alpha_0 + \alpha_1 D_i + \beta_0 x_i + \beta_1 (D_i x_i) + e_i,$$

which yields two regression lines with different intercepts and slopes. The slope of the "male" regression line is mow $\beta_0 + \beta_1$, whereas the slope of the "female" regression line is β_0. By testing $\beta_1 = 0$, we can check whether teaching experience is treated the same in determining salaries for male and female teachers.

Suppose that we want to know if the education level of the head of household affects family consumption pattern. We may classify data into three groups: below high school, high school only, college or higher. Let $D_{1i} = 1$ if i has a high school degree only and $D_{1i} = 0$ otherwise, and $D_{2i} = 1$ if i has a college or higher degree and $D_{2i} = 0$ otherwise. Then, similar to the previous example, the following specification,

$$y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \beta x_i + e_i,$$

yields three regression lines. The below-high-school regression line has the intercept α_0, the high-school regression line has the intercept $\alpha_0 + \alpha_1$, and the college regression line has the intercept $\alpha_0 + \alpha_2$. Various interesting hypotheses can be tested based on this specification.

Remark: The preceding examples show that, when a specification contains a constant term, the number of dummy variables is always one *less* than the number of categories that dummy variables try to classify. Otherwise, the specification has exact multicollinearity; this is the so-called "dummy variable trap."

3.7 Limitations of the Classical Conditions

The previous estimation and testing results are based on the classical conditions. As these conditions may be violated in practice, it is important to understand their limitations.

Condition [A1] postulates that explanatory variables are non-stochastic. Although this condition is quite convenient and facilitates our analysis, it is not practical. When the dependent variable and regressors are economic variables, it does not make too much sense to treat only the dependent variable as a random variable. This condition may also be violated when a lagged dependent variable is included as a regressor, as in many time-series analysis. Hence, it would be more reasonable to allow regressors to be random as well.

In [A2](i), the linear specification $\boldsymbol{X}\boldsymbol{\beta}$ is assumed to be correct up to some unknown parameters. It is possible that the systematic component $\mathbb{E}(\boldsymbol{y})$ is in fact a non-linear function of $\boldsymbol{X}$. If so, the estimated regression hyperplane could be very misleading. For example, an economic relation may change from one regime to another at some time point so that $\mathbb{E}(\boldsymbol{y})$ is better characterized by a piecewise liner function. This is known as the problem of *structural change*; see e.g., Exercise 3.14. Even when $\mathbb{E}(\boldsymbol{y})$ is a linear function, the specified $\boldsymbol{X}$ may include some irrelevant variables or omit some important variables. Example 3.6 shows that in the former case, the OLS estimator $\hat{\boldsymbol{\beta}}_T$ remains unbiased but is less efficient. In the latter case, it can be shown that $\hat{\boldsymbol{\beta}}_T$ is biased but with a smaller variance-covariance matrix; see Exercise 3.6.

Condition [A2](ii) may also easily break down in many applications. For example, when y_t is the consumption of the t^{th} household, it is likely that y_t has smaller variation for low-income families than for high-income families. When y_t denotes the GDP growth rate of the t^{th} year, it is also likely that y_t are correlated over time. In both cases, the variance-covariance matrix of $\boldsymbol{y}$ cannot be expressed as $\sigma_o^2\boldsymbol{I}_T$. A consequence of the failure of [A2](ii) is that the OLS estimator for $\text{var}(\hat{\boldsymbol{\beta}}_T)$, $\hat{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, is biased, which in turn renders the tests discussed in Section 3.4 invalid.

Condition [A3] may fail when y_t have non-normal distributions. Although the BLUE property of the OLS estimator does not depend on normality, [A3] is crucial for deriving the distribution results in Section 3.4. When [A3] is not satisfied, the usual t and F tests do not have the desired t and F distributions, and their exact distributions are typically unknown. This causes serious problems for hypothesis testing.

Our discussion thus far suggests that the classical conditions are quite restrictive.

In subsequent chapters, we will try to relax these conditions and discuss more generally applicable methods. These methods play an important role in contemporary empirical studies.

Exercises

3.1 Construct a linear regression model for each equation below:

$$y = \alpha\, x^{\beta}, \qquad y = \alpha\, e^{\beta x}, \qquad y = \frac{x}{\alpha x - \beta}, \qquad y = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}.$$

3.2 Use the general formula (3.4) to find the OLS estimators from the specifications below:

$$y_t = \alpha + \beta x_t + e,\ t = 1, \ldots, T,$$

$$y_t = \alpha + \beta(x_t - \bar{x}) + e,\ t = 1, \ldots, T,$$

$$y_t = \beta x_t + e,\ t = 1, \ldots, T.$$

Compare the resulting regression lines.

3.3 Given the specification $y_t = \alpha + \beta x_t + e$, $t = 1, \ldots, T$, assume that the classical conditions hold. Let $\hat{\alpha}_T$ and $\hat{\beta}_T$ be the OLS estimators for α and β, respectively.

(a) Apply the general formula of Theorem 3.4(c) to show that

$$\begin{aligned}
\operatorname{var}(\hat{\alpha}_T) &= \sigma_o^2 \frac{\sum_{t=1}^T x_t^2}{T \sum_{t=1}^T (x_t - \bar{x})^2}, \\
\operatorname{var}(\hat{\beta}_T) &= \sigma_o^2 \frac{1}{\sum_{t=1}^T (x_t - \bar{x})^2}, \\
\operatorname{cov}(\hat{\alpha}_T, \hat{\beta}_T) &= -\sigma_o^2 \frac{\bar{x}}{\sum_{t=1}^T (x_t - \bar{x})^2}.
\end{aligned}$$

What kind of data can make the variances of the OLS estimators smaller?

(b) Suppose that a prediction $\hat{y}_{T+1} = \hat{\alpha}_T + \hat{\beta}_T x_{T+1}$ is made based on the new observation x_{T+1}. Show that

$$\begin{aligned}
\mathbb{E}(\hat{y}_{T+1} - y_{T+1}) &= 0, \\
\operatorname{var}(\hat{y}_{T+1} - y_{T+1}) &= \sigma_o^2 \left(1 + \frac{1}{T} + \frac{(x_{T+1} - \bar{x})^2}{\sum_{t=1}^T (x_t - \bar{x})^2} \right).
\end{aligned}$$

What kind of x_{T+1} can make the variance of prediction error smaller?

3.4 Given the specification (3.1), suppose that $\boldsymbol{X}$ is not of full column rank. Does there exist a unique $\hat{\boldsymbol{y}} \in \text{span}(\boldsymbol{X})$ that minimizes $(\boldsymbol{y}-\hat{\boldsymbol{y}})'(\boldsymbol{y}-\hat{\boldsymbol{y}})$? If yes, is there a unique $\hat{\boldsymbol{\beta}}_T$ such that $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_T$? Why or why not?

3.5 Given the estimated model

$$y_t = \hat{\beta}_{1,T} + \hat{\beta}_{2,T} x_{t2} + \cdots + \hat{\beta}_{k,T} x_{tk} + \hat{e}_t,$$

consider the *standardized* regression:

$$y_t^* = \hat{\beta}_{2,T}^* x_{t2}^* + \cdots + \hat{\beta}_{k,T}^* x_{tk}^* + \hat{e}_t^*,$$

where $\hat{\beta}_{i,T}^*$ are known as the *beta coefficients*, and

$$y_t^* = \frac{y_t - \bar{y}}{s_y}, \qquad x_{ti}^* = \frac{x_{ti} - \bar{x}_i}{s_{x_i}}, \qquad \hat{\epsilon}_t^* = \frac{\hat{e}_t}{s_y},$$

with $s_y^2 = (T-1)^{-1}\sum_{t=1}^T (y_t - \bar{y})^2$ is the sample variance of y_t and for each i, $s_{x_i}^2 = (T-1)^{-1}\sum_{t=1}^T (x_{ti} - \bar{x}_i)^2$ is the sample variance of x_{ti}. What is the relationship between $\hat{\beta}_{i,T}^*$ and $\hat{\beta}_{i,T}$? Give an interpretation of the beta coefficients.

3.6 Given the following specification

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + e,$$

where $\boldsymbol{X}_1$ $(T \times k_1)$ is a non-stochastic matrix, let $\hat{\boldsymbol{b}}_{1,T}$ denote the resulting OLS estimator. Suppose that $\mathbb{E}(\boldsymbol{y}) = \boldsymbol{X}_1\boldsymbol{b}_1 + \boldsymbol{X}_2\boldsymbol{b}_2$ for some $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$, where $\boldsymbol{X}_2$ $(T \times k_2)$ is also a non-stochastic matrix and $\boldsymbol{b}_2 \neq \boldsymbol{0}$.

(a) Is $\hat{\boldsymbol{b}}_{1,T}$ unbiased?

(b) Is $\hat{\sigma}_T^2$ unbiased?

(c) What is $\text{var}(\hat{\boldsymbol{b}}_{1,T})$?

(d) Let $\hat{\boldsymbol{\beta}}_T = (\hat{\boldsymbol{\beta}}_{1,T}' \ \hat{\boldsymbol{\beta}}_{2,T}')'$ denote the OLS estimator obtained from estimating the specification: $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}$. Compare $\text{var}(\hat{\boldsymbol{\beta}}_{1,T})$ and $\text{var}(\hat{\boldsymbol{b}}_T)$.

(e) Does your result in (d) change when $\boldsymbol{X}_1'\boldsymbol{X}_2 = \boldsymbol{0}$?

3.7 Given the specification (3.1), will the changes below affect the resulting OLS estimator $\hat{\boldsymbol{\beta}}_T$, t ratios, and R^2?

(a) $\boldsymbol{y}^* = 1000 \times \boldsymbol{y}$ and $\boldsymbol{X}$ are used as the dependent and explanatory variables.

(b) $\boldsymbol{y}$ and $\boldsymbol{X}^* = 1000 \times \boldsymbol{X}$ are used as the dependent and explanatory variables.

(c) $\boldsymbol{y}^*$ and $\boldsymbol{X}^*$ are used as the dependent and explanatory variables.

3.8 Let R_k^2 denote the centered R^2 obtained from the model with k explanatory variables.

(a) Show that

$$R_k^2 = \sum_{i=1}^{k} \hat{\beta}_{iT} \frac{\sum_{t=1}^{T}(x_{ti} - \bar{x}_i)y_t}{\sum_{t=1}^{T}(y_t - \bar{y})^2},$$

where $\hat{\beta}_{iT}$ is the ith element of $\hat{\boldsymbol{\beta}}_T$, $\bar{x}_i = \sum_{t=1}^{T} x_{ti}/T$, and $\bar{y} = \sum_{t=1}^{T} y_t/T$.

(b) Show that $R_k^2 \geq R_{k-1}^2$.

3.9 Consider the following two regression lines: $\hat{y} = \hat{\alpha} + \hat{\beta}x$ and $\hat{x} = \hat{\gamma} + \hat{\delta}y$. At which point do these two lines intersect? Using the result in Exercise 3.8 to show that these two regression lines coincide if and only if the centered R^2s for both regressions are one.

3.10 Given the specification (3.1), suppose that $\boldsymbol{X}$ does *not* contain the constant term. Show that the centered R^2 need not be bounded between zero and one if it is computed as (3.7).

3.11 Rearrange the matrix $\boldsymbol{X}$ as $[\boldsymbol{x}_i \ \boldsymbol{X}_i]$, where $\boldsymbol{x}_i$ is the ith column of $\boldsymbol{X}$. Let $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ denote the residual vectors of regressing $\boldsymbol{y}$ on $\boldsymbol{X}_i$ and $\boldsymbol{x}_i$ on $\boldsymbol{X}_i$, respectively. Define the *partial correlation coefficient* of $\boldsymbol{y}$ and $\boldsymbol{x}_i$ as

$$r_i = \frac{\boldsymbol{u}_i'\boldsymbol{v}_i}{(\boldsymbol{u}_i'\boldsymbol{u}_i)^{1/2}(\boldsymbol{v}_i'\boldsymbol{v}_i)^{1/2}}.$$

Let R_i^2 and R^2 be obtained from the regressions of $\boldsymbol{y}$ on $\boldsymbol{X}_i$ and $\boldsymbol{y}$ on $\boldsymbol{X}$, respectively.

(a) Apply the Frisch-Waugh-Lovell Theorem to show

$$\boldsymbol{I} - \boldsymbol{P} = (\boldsymbol{I} - \boldsymbol{P}_i) - \frac{(\boldsymbol{I} - \boldsymbol{P}_i)\boldsymbol{x}_i\boldsymbol{x}_i'(\boldsymbol{I} - \boldsymbol{P}_i)}{\boldsymbol{x}_i'(\boldsymbol{I} - \boldsymbol{P}_i)\boldsymbol{x}_i},$$

where $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ and $\boldsymbol{P}_i = \boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'$. Note that this result can also be derived using the matrix inversion formula (e.g. Greene (1993, p. 27)).

(b) Show that $(1 - R^2)/(1 - R_i^2) = 1 - r_i^2$, and use this result to verify

$$R^2 - R_i^2 = r_i^2(1 - R_i^2).$$

What does this result tell you?

(c) Let τ_i denote the t ratio of $\hat{\beta}_{iT}$, the ith element of $\hat{\boldsymbol{\beta}}_T$ obtained from regressing $\boldsymbol{y}$ on $\boldsymbol{X}$. First show that $\tau_i^2 = (T-k)r_i^2/(1-r_i^2)$, and use this result to verify

$$r_i^2 = \tau_i^2/(\tau_i^2 + T - k).$$

(d) Combine the results in (b) and (c) to show

$$R^2 - R_i^2 = \tau_i^2(1 - R^2)/(T - k).$$

What does this result tell you?

3.12 Suppose that a linear model with k explanatory variables has been estimated.

(a) Show that $\hat{\sigma}_T^2 = \text{Centered TSS}(1 - \bar{R}^2)/(T-1)$. What does this result tell you?

(b) Suppose that we want to test the hypothesis that s coefficients are zero. Show that the F statistic can be written as

$$\varphi = \frac{(T-k+s)\hat{\sigma}_{\text{c}}^2 - (T-k)\hat{\sigma}_{\text{u}}^2}{s\hat{\sigma}_{\text{u}}^2},$$

where $\hat{\sigma}_{\text{c}}^2$ and $\hat{\sigma}_{\text{u}}^2$ are the variance estimates of the constrained and unconstrained models, respectively. Let $a = (T-k)/s$. Show that

$$\frac{\hat{\sigma}_{\text{c}}^2}{\hat{\sigma}_{\text{u}}^2} = \frac{a + \varphi}{a + 1}.$$

(c) Based on the results in (a) and (b), what can you say when $\varphi > 1$ and $\varphi < 1$?

3.13 For the linear specification $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, an alternative expression of $k-m$ linear restrictions on $\boldsymbol{\beta}$ can be expressed as $\boldsymbol{\beta} = \boldsymbol{S}\boldsymbol{\theta} + \boldsymbol{d}$, where $\boldsymbol{\theta}$ is a m-dimensional vector of unknown parameters, $\boldsymbol{S}$ is a $k \times m$ matrix of pre-specified constants with full column rank, and $\boldsymbol{c}$ is a vector of pre-specified constants.

(a) By incorporating this restriction into the specification, find the OLS estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$.

(b) The *constrained least squares estimator* of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}_c = \boldsymbol{S}\hat{\theta} + \boldsymbol{d}$. Show that

$$\hat{\boldsymbol{\beta}}_c = \boldsymbol{Q}_S\hat{\boldsymbol{\beta}} + (\boldsymbol{I} - \boldsymbol{Q}_S)\boldsymbol{d},$$

where $\boldsymbol{Q}_S = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{X}'\boldsymbol{X}$. Is this decomposition orthogonal?

(c) Show that

$$\boldsymbol{X}\hat{\boldsymbol{\beta}}_c = \boldsymbol{P}_{XS}\boldsymbol{y} + (\boldsymbol{I} - \boldsymbol{P}_{XS})\boldsymbol{X}\boldsymbol{d},$$

where $\boldsymbol{P}_{XS} = \boldsymbol{X}\boldsymbol{S}(\boldsymbol{S}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{X}'$. Use a graph to illustrate this result.

3.14 (The Chow Test) Consider the model of a one-time structural change at a known change point:

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{0} \\ \boldsymbol{X}_2 & \boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_o \\ \boldsymbol{\delta}_o \end{bmatrix} + \begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \end{bmatrix},$$

where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are $T_1 \times 1$ and $T_2 \times 1$, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are $T_1 \times k$ and $T_2 \times k$, respectively. The null hypothesis is $\boldsymbol{\delta}_o = \boldsymbol{0}$. How would you test this hypothesis based on the constrained and unconstrained models?

References

Davidson, Russell and James G. MacKinnon (1993). *Estimation and Inference in Econometrics*, New York, NY: Oxford University Press.

Goldberger, Arthur S. (1991). *A Course in Econometrics*, Cambridge, MA: Harvard University Press.

Greene, William H. (2000). *Econometric Analysis*, 4th ed., Upper Saddle River, NJ: Prentice Hall.

Harvey, Andrew C. (1990). *The Econometric Analysis of Time Series*, Second edition., Cambridge, MA: MIT Press.

Intriligator, Michael D., Ronald G. Bodkin, and Cheng Hsiao (1996). *Econometric Models, Techniques, and Applications*, Second edition, Upper Saddle River, NJ: Prentice Hall.

Johnston, J. (1984). *Econometric Methods*, Third edition, New York, NY: McGraw-Hill.

Judge, Georgge G., R. Carter Hill, William E. Griffiths, Helmut Lütkepohl, and Tsoung-Chao Lee (1988). *Introduction to the Theory and Practice of Econometrics*, Second edition, New York, NY: Wiley.

Maddala, G. S. (1992). *Introduction to Econometrics*, Second edition, New York, NY: Macmillan.

Manski, Charles F. (1991). Regression, *Journal of Economic Literature*, **29**, 34–50.

Rao, C. Radhakrishna (1973). *Linear Statistical Inference and Its Applications*, Second edition, New York, NY: Wiley.

Ruud, Paul A. (2000). *An Introduction to Classical Econometric Theory*, New York, NY: Oxford University Press.

Theil, Henri (1971). *Principles of Econometrics*, New York, NY: Wiley.

Chapter 4

Generalized Least Squares Theory

4.1 Introduction

In Chapter 3.7 we have seen that the classical conditions need not hold in practice. Although these conditions have no effect on the OLS method per se, they do affect the properties of the OLS estimators and resulting test statistics. In particular, when the elements of $\boldsymbol{y}$ have unequal variances and/or are correlated, there is no guarantee that the OLS estimator is the most efficient within the class of linear unbiased (or the class of unbiased) estimators. Moreover, hypothesis testing based on the standard OLS estimator of the variance-covariance matrix becomes invalid.

In this chapter, the method of *generalized least squares* (GLS) is introduced to improve upon estimation efficiency. A drawback of the GLS method is that it is difficult to implement. In practice, certain structures (assumptions) must be imposed on var($\boldsymbol{y}$) so that a feasible GLS estimator can be computed. This approach results in two further difficulties, however. First, the postulated structures on var($\boldsymbol{y}$) need not be correctly specified. Consequently, the resulting feasible GLS estimator may not be as efficient as one would like. Second, the finite-sample properties of feasible GLS estimators are not easy to establish. Exact tests based on the feasible GLS results are thus not readily available. More detailed discussions of the GLS theory can also be found in e.g., Amemiya (1985) and Greene (2000).

4.2 The Method of Generalized Least Squares

4.2.1 When y Does Not Have a Scalar Covariance Matrix

Given the linear specification (3.1):

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},$$

suppose that, in addition to the conditions [A1] and [A2](i),

$$\mathrm{var}(\boldsymbol{y}) = \boldsymbol{\Sigma}_o,$$

where $\boldsymbol{\Sigma}_o$ is a positive definite matrix but cannot be written as $\sigma_o^2\boldsymbol{I}_T$ for any positive number σ_o^2. That is, the elements of $\boldsymbol{y}$ may not have a constant variance, nor are they required to be uncorrelated. As [A1] and [A2](i) still hold, the OLS estimator $\hat{\boldsymbol{\beta}}_T$ remains unbiased by Theorem 3.4(a), and

$$\mathrm{var}(\hat{\boldsymbol{\beta}}_T) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_o\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}, \tag{4.1}$$

in contrast with Theorem 3.4(c). In view of Theorem 3.5, there is no guarantee that the OLS estimator is the BLUE for $\boldsymbol{\beta}_o$. Similarly, when [A3] fails such that

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}_o, \boldsymbol{\Sigma}_o),$$

we have

$$\hat{\boldsymbol{\beta}}_T \sim N(\boldsymbol{\beta}_o,\ (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_o\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1});$$

cf. Theorem 3.7(a). In this case, $\hat{\boldsymbol{\beta}}_T$ need not be the BUE for $\boldsymbol{\beta}_o$.

Apart from efficiency, a more serious consequence of the failure of [A3] is that the statistical tests based on the standard OLS estimation results become invalid. Recall that the OLS estimator for $\mathrm{var}(\hat{\boldsymbol{\beta}}_T)$ is

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}_T) = \hat{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X})^{-1},$$

which is, in general, a biased estimator for (4.1). As the t and F statistics depend on the elements of $\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}_T)$, they no longer have the desired t and F distributions under the null hypothesis. Consequently, the inferences based on these tests become invalid.

4.2.2 The GLS Estimator

The GLS method focuses on the efficiency issue resulted from the failure of the classical condition [A2](ii). Let $\boldsymbol{G}$ be a $T \times T$ non-stochastic matrix. Consider the "transformed" specification

$$\boldsymbol{G}\boldsymbol{y} = \boldsymbol{G}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{G}\boldsymbol{e},$$

where $\boldsymbol{G}\boldsymbol{y}$ denotes the transformed dependent variable and $\boldsymbol{G}\boldsymbol{X}$ is the matrix of transformed explanatory variables. It can be seen that $\boldsymbol{G}\boldsymbol{X}$ also has full column rank k provided that $\boldsymbol{G}$ is nonsingular. Thus, the identification requirement for the specification (3.1) carries over under nonsingular transformations. It follows that $\boldsymbol{\beta}$ can still be estimated by the following OLS estimator:

$$(\boldsymbol{X}'\boldsymbol{G}'\boldsymbol{G}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{G}'\boldsymbol{G}\boldsymbol{y}. \tag{4.2}$$

Given that the original variables $\boldsymbol{y}$ and $\boldsymbol{X}$ satisfy [A1] and [A2](i), it is easily seen that the transformed variables $\boldsymbol{G}\boldsymbol{y}$ and $\boldsymbol{G}\boldsymbol{X}$ also satisfy these two conditions because $\boldsymbol{G}\boldsymbol{X}$ is non-stochastic and $\mathbb{E}(\boldsymbol{G}\boldsymbol{y}) = \boldsymbol{G}\boldsymbol{X}\boldsymbol{\beta}_o$. When $\mathrm{var}(\boldsymbol{y}) = \boldsymbol{\Sigma}_o$,

$$\mathrm{var}(\boldsymbol{G}\boldsymbol{y}) = \boldsymbol{G}\boldsymbol{\Sigma}_o\boldsymbol{G}'.$$

If $\boldsymbol{G}$ is such that $\boldsymbol{G}\boldsymbol{\Sigma}_o\boldsymbol{G}' = \sigma_o^2\boldsymbol{I}_T$ for some positive number σ_o^2, the condition [A2](ii) would also hold. Since the classical conditions are all satisfied, the OLS estimator

$$(\boldsymbol{X}'\boldsymbol{G}'\boldsymbol{G}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{G}'\boldsymbol{G}\boldsymbol{y}$$

is still the BLUE for $\boldsymbol{\beta}_o$ by Theorem 3.5. This suggests that, as far as efficiency is concerned, one should estimate $\boldsymbol{\beta}$ from the transformed specification such that the transformation matrix G is nonsingular and $\boldsymbol{G}\boldsymbol{\Sigma}_o\boldsymbol{G}' = \sigma_o^2\boldsymbol{I}_T$.

To find a desirable transformation matrix G, note that $\boldsymbol{\Sigma}_o$ is symmetric and positive definite and that $\boldsymbol{\Sigma}_o$ can be orthogonally diagonalized as $\boldsymbol{C}'\boldsymbol{\Sigma}_o\boldsymbol{C} = \boldsymbol{\Lambda}$, where $\boldsymbol{C}$ is the matrix of eigenvectors corresponding to the matrix of eigenvalues $\boldsymbol{\Lambda}$. For $\boldsymbol{\Sigma}_o^{-1/2} = \boldsymbol{C}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{C}'$ (or $\boldsymbol{\Sigma}_o^{-1/2} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{C}'$), we have

$$\boldsymbol{\Sigma}_o^{-1/2}\boldsymbol{\Sigma}_o\boldsymbol{\Sigma}_o^{-1/2\prime} = \boldsymbol{I}_T.$$

This result immediately suggests that the desired matrix G should be proportional to $\boldsymbol{\Sigma}_o^{-1/2}$, i.e., $\boldsymbol{G} = c\,\boldsymbol{\Sigma}_o^{-1/2}$ for some constant c. Given this choice of G, we have

$$\mathrm{var}(\boldsymbol{G}\boldsymbol{y}) = \boldsymbol{G}\boldsymbol{\Sigma}_o\boldsymbol{G}' = c^2\boldsymbol{I}_T,$$

a scalar covariance matrix, so that [A2](ii) also holds. It follows that the estimator (4.2) with $\boldsymbol{G} = c\,\boldsymbol{\Sigma}_o^{-1/2}$ is the BLUE for $\boldsymbol{\beta}_o$. This estimator is known as the GLS estimator and reads

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = (c^2\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}(c^2\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{y}) = (\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{y}. \tag{4.3}$$

It should be clear that the GLS estimator cannot be computed unless $\boldsymbol{\Sigma}_o$ is known. As $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ does not depend on c, it suffices to choose the transformation matrix $\boldsymbol{G} = \boldsymbol{\Sigma}_o^{-1/2}$.

For $\boldsymbol{G} = \boldsymbol{\Sigma}_o^{-1/2}$, let $\boldsymbol{y}^* = \boldsymbol{G}\boldsymbol{y}$, $\boldsymbol{X}^* = \boldsymbol{G}\boldsymbol{X}$, and $e^* = \boldsymbol{G}\boldsymbol{e}$. The transformed specification is

$$\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{e}^*, \tag{4.4}$$

As the GLS estimator is the OLS estimator for the specification (4.4), it can also be interpreted as a minimizer of the following GLS criterion function:

$$Q(\boldsymbol{\beta}; \boldsymbol{\Sigma}_o) = \frac{1}{T}(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta})'(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}) = \frac{1}{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{4.5}$$

This is a generalized version of the standard OLS criterion function (3.2) because it is the average of a *weighted* sum of squared errors. Thus, the GLS and OLS estimators are not equivalent in general, except in some exceptional cases; see e.g. Exercise 4.1.

Similar to the OLS method, define the vector of GLS fitted values as

$$\hat{\boldsymbol{y}}_{\text{GLS}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{y}.$$

The vector of GLS residuals is

$$\hat{\boldsymbol{e}}_{\text{GLS}} = \boldsymbol{y} - \hat{\boldsymbol{y}}_{\text{GLS}}.$$

As $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}$ is idempotent but not symmetric, $\hat{\boldsymbol{y}}_{\text{GLS}}$ is an *oblique* (but not orthogonal) projection of $\boldsymbol{y}$ onto span($\boldsymbol{X}$). It can also be verified that the vector of GLS residuals is not orthogonal to $\boldsymbol{X}$ or any linear combination of the column vectors of $\boldsymbol{X}$, i.e.,

$$\hat{\boldsymbol{e}}_{\text{GLS}}'\boldsymbol{X} = \boldsymbol{y}'[\boldsymbol{I}_T - \boldsymbol{\Sigma}_o^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{X} \neq \boldsymbol{0}.$$

In fact, $\hat{\boldsymbol{e}}_{\text{GLS}}$ is orthogonal to span($\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X}$). It follows that

$$\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} \leq \hat{\boldsymbol{e}}_{\text{GLS}}'\hat{\boldsymbol{e}}_{\text{GLS}}.$$

This shows that the OLS method still yields a better fit of original data.

Remark: A natural measure of goodness of fit for the specification (3.1) estimated using the GLS method is

$$\text{Centered } R^2_{\text{GLS}} = 1 - \frac{\hat{\boldsymbol{e}}'_{\text{GLS}}\hat{\boldsymbol{e}}_{\text{GLS}}}{\text{Centered TSS of } \boldsymbol{y}},$$

where the denominator is the TSS of the original dependent variable y. A major problem of this measure is that it need not be bounded between zero and one; see Exercise 4.2. Thus, R^2_{GLS} is not a proper criterion for model comparison. Using R^2 from the transformed specification (4.4) is also inadequate because it can only measure the variation of the transformed dependent variable y^*, but not the variation of the original variable y.

4.2.3 Properties of the GLS Estimator

We have seen that the GLS estimator is, by construction, the BLUE for $\boldsymbol{\beta}_o$ under [A1] and [A2](i). Its variance-covariance matrix is

$$\text{var}(\hat{\boldsymbol{\beta}}_{\text{GLS}}) = \text{var}((\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{y}) = (\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}. \tag{4.6}$$

These results are summarized below.

Theorem 4.1 (Aitken) *Given the specification (3.1), suppose that* [A1] and [A2](i) *hold and that* $\text{var}(\boldsymbol{y}) = \boldsymbol{\Sigma}_o$ *is a positive definite matrix. Then* $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ *is the BLUE for* $\boldsymbol{\beta}_o$ *with the variance-covariance matrix* $(\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}$.

As the GLS estimator is the BLUE,

$$\text{var}(\hat{\boldsymbol{\beta}}_T) - \text{var}(\hat{\boldsymbol{\beta}}_{\text{GLS}})$$

must be a positive semi-definite matrix. This can also be verified directly; see Exercise 4.3.

For convenience, we introduce the following condition.

[A3′] $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}_o, \boldsymbol{\Sigma}_o)$, where $\boldsymbol{\Sigma}_o$ is a positive definite matrix.

The following result is an immediate consequence of Theorem 3.7(a).

Theorem 4.2 *Given the specification (3.1), suppose that* [A1] *and* [A3′] *hold. Then*

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} \sim N(\boldsymbol{\beta}_o, (\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}).$$

Moreover, if we believe that [A3′] is true, the log-likelihood function is

$$\log L(\boldsymbol{\beta}; \boldsymbol{\Sigma}_o) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\log(\det(\boldsymbol{\Sigma}_o)) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \quad (4.7)$$

The first order conditions of maximizing this log-likelihood function with respect to $\boldsymbol{\beta}$ are

$$\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0},$$

so that the MLE is

$$\tilde{\boldsymbol{\beta}}_T = (\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{y}.$$

Thus, when $\boldsymbol{\Sigma}_o$ is known, the GLS estimator is also the MLE under [A3′]. The information matrix is then

$$\mathbb{E}[\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X}]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_o} = \boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X}.$$

The variance-covariance matrix of the GLS estimator thus also achieves the Crámer-Rao lower bound. We have shown:

Theorem 4.3 *Given the specification (3.1), suppose that* [A1] *and* [A3′] *hold. Then* $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ *is the BUE for* $\boldsymbol{\beta}_o$.

Under the null hypothesis $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r}$, it is readily seen from Theorem 4.2 that

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}}_{\text{GLS}} - \boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{\Sigma}_o^{-1}\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_{\text{GLS}} - \boldsymbol{r}) \sim \chi^2(q).$$

The left-hand side can serve as a test statistic provided that $\boldsymbol{\Sigma}_o$ is known.

4.2.4 FGLS Estimator

In practice, $\boldsymbol{\Sigma}_o$ is typically unknown and must be estimated. Substituting an estimator $\hat{\boldsymbol{\Sigma}}_T$ for $\boldsymbol{\Sigma}_o$ in (4.3) yields the *feasible generalized least squares* (FGLS) estimator

$$\hat{\boldsymbol{\beta}}_{\text{FGLS}} = (\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_T^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\hat{\boldsymbol{\Sigma}}_T^{-1}\boldsymbol{y}.$$

which is readily computed from data. Note, however, that $\boldsymbol{\Sigma}_o$ contains too many $(T(T+1)/2)$ parameters. Proper estimation of $\boldsymbol{\Sigma}_o$ would not be possible unless further restrictions on $\boldsymbol{\Sigma}_o$ are imposed. Under different assumptions on $\text{var}(\boldsymbol{y})$, $\boldsymbol{\Sigma}_o$ has a simpler structure with much fewer (say, $p \ll T$) unknown parameters and may be

properly estimated; see Sections 4.3 and 4.4. FGLS estimation crucially depends on these assumptions.

A clear disadvantage of FGLS estimators is that their finite sample properties are usually unknown. Note that $\hat{\boldsymbol{\Sigma}}_T$ is, in general, a function of $\boldsymbol{y}$, so that $\hat{\boldsymbol{\beta}}_{\text{FGLS}}$ is a complex function of the elements of $\boldsymbol{y}$. It is therefore difficult, if not impossible, to derive the finite-sample properties, such as expectation, variance and distribution, of $\hat{\boldsymbol{\beta}}_{\text{FGLS}}$. Consequently, the efficiency gain of an FGLS estimator is not at all clear, and exact tests are not available. One must rely on the asymptotic properties of $\hat{\boldsymbol{\beta}}_{\text{FGLS}}$ to draw statistical inferences.

4.3 Heteroskedasticity

In this section, we consider a simpler structure of $\boldsymbol{\Sigma}_o$ such that $\boldsymbol{\Sigma}_o$ is diagonal with possibly different diagonal elements:

$$\boldsymbol{\Sigma}_o = \text{diag}[\sigma_1^2, \ldots, \sigma_T^2] = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_T^2 \end{bmatrix}, \tag{4.8}$$

where diag is the operator that puts its arguments on the main diagonal of a matrix. That is, the elements of $\boldsymbol{y}$ are uncorrelated but may have different variances. When y_t, $t = 1, \ldots, T$, have a constant variance, they are said to be *homoskedastic*; otherwise, they are *heteroskedastic*.

To compute the GLS estimator, the desired transformation matrix is

$$\boldsymbol{\Sigma}_o^{-1/2} = \begin{bmatrix} \sigma_1^{-1} & 0 & \cdots & 0 \\ 0 & \sigma_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_T^{-1} \end{bmatrix}.$$

As $\boldsymbol{\Sigma}_o$ still contains T unknown parameters, an even simpler structure of $\boldsymbol{\Sigma}_o$ is needed to ensure proper FGLS estimation.

4.3.1 Tests for Heteroskedasticity

It is clear that the OLS method would prevail unless there is evidence that $\boldsymbol{\Sigma}_o \neq \sigma_o^2 \boldsymbol{I}_T$. It is therefore natural to first study the tests of the null hypothesis of *homoskedasticity*

against some form of *heteroskedasticity*. Such tests are usually based on some simplified parametric specifications of $\mathrm{var}(y_t)$.

The simplest possible form of heteroskedastic y_t is *groupwise heteroskedasticity*. Suppose that data can be classified into two groups: group one contains T_1 observations with the constant variance σ_1^2, and group two contains T_2 observations with the constant variance σ_2^2. This assumption simplifies $\mathbf{\Sigma}_o$ in (4.8) to a matrix of only two unknown parameters:

$$\mathbf{\Sigma}_o = \begin{bmatrix} \sigma_1^2 \boldsymbol{I}_{T_1} & \mathbf{0} \\ \mathbf{0} & \sigma_2^2 \boldsymbol{I}_{T_2} \end{bmatrix}, \tag{4.9}$$

The null hypothesis of homoskedasticity is $\sigma_1^2 = \sigma_2^2 = \sigma_o^2$; the alternative hypothesis is, without loss of generality, $\sigma_1^2 > \sigma_2^2$.

Consider now two regressions based on the observations of the group one and group two, respectively. Let $\hat{\sigma}_{T_1}^2$ and $\hat{\sigma}_{T_2}^2$ denote the resulting OLS variance estimates. Intuitively, whether $\hat{\sigma}_{T_1}^2$ is "close" to $\hat{\sigma}_{T_2}^2$ constitutes an evidence for or against the null hypothesis. Under [A1] and [A3′] with (4.9),

$$(T_1 - k)\hat{\sigma}_{T_1}^2/\sigma_1^2 \sim \chi^2(T_1 - k),$$
$$(T_2 - k)\hat{\sigma}_{T_2}^2/\sigma_2^2 \sim \chi^2(T_2 - k),$$

by Theorem 3.7(b). As y_t are independent, these two χ^2 random variables are also mutually independent. Note that $\hat{\sigma}_{T_1}^2$ and $\hat{\sigma}_{T_2}^2$ must be computed from separate regressions so as to ensure independence. Then under the null hypothesis,

$$\varphi := \frac{\hat{\sigma}_{T_1}^2}{\hat{\sigma}_{T_2}^2} = \frac{(T_1 - k)\hat{\sigma}_{T_1}^2}{\sigma_o^2(T_1 - k)} \Bigg/ \frac{(T_2 - k)\hat{\sigma}_{T_2}^2}{\sigma_o^2(T_2 - k)} \sim F(T_1 - k, T_2 - k);$$

this is the F test for groupwise heteroskedasticity.

More generally, the variances of y_t may be changing with the values of a particular explanatory variable, say x_j. That is, for some constant $c > 0$,

$$\sigma_t^2 = c\, x_{tj}^2.$$

Thus, the larger the magnitude of x_{tj}, the greater is σ_t^2. An interesting feature of this specification is that σ_t^2 may take distinct values for every t, yet $\mathbf{\Sigma}_o$ contains only one unknown parameter c. The null hypothesis is then $\sigma_t^2 = \sigma_o^2$ for all t, and the alternative hypothesis is, without loss of generality,

$$\sigma_{(1)}^2 \geq \sigma_{(2)}^2 \geq \ldots \sigma_{(T)}^2,$$

where $\sigma^2_{(i)}$ denotes the i^{th} largest variance. The so-called Goldfeld-Quandt test is of the same form as the F test for groupwise heteroskedasticity but with the following data grouping procedure.

(1) Rearrange observations according to the values of some explanatory variable x_j in a descending order.

(2) Divide the rearranged data set into three groups with T_1, T_m, and T_2 observations, respectively.

(3) Drop the T_m observations in the middle group and perform separate OLS regressions using the data in the first and third groups.

(4) The statistic is the ratio of the variance estimates:

$$\hat{\sigma}^2_{T_1}/\hat{\sigma}^2_{T_2} \sim F(T_1 - k, T_2 - k).$$

If the data are rearranged according to the values of x_j in an ascending order, the resulting statistic should be computed as

$$\hat{\sigma}^2_{T_2}/\hat{\sigma}^2_{T_1} \sim F(T_2 - k, T_1 - k).$$

In a time-series study, the variances may be decreasing (increasing) over time. In this case, data rearrangement would not be needed. Note that dropping the observations in the middle group enhances the test's ability of discriminating variances in the first and third groups. It is usually suggested that no more than one third of the observations should be dropped; it is also typical to set $T_1 \approx T_2$. Clearly, this test would be powerful provided that one can correctly identify the source of heteroskedasticity (i.e., the explanatory variable that determines variances). On the other hand, finding such an explanatory variable may not be easy.

An even more general form of heteroskedastic covariance matrix is such that the diagonal elements

$$\sigma^2_t = h(\alpha_0 + \boldsymbol{z}_t'\boldsymbol{\alpha}_1),$$

where h is some function and $\boldsymbol{z}_t$ is a $p \times 1$ vector of exogenous variables affecting the variances of y_t. This assumption simplifies $\boldsymbol{\Sigma}_o$ to a matrix of $p+1$ unknown parameters. Tests against this class of alternatives can be derived under the likelihood framework, and their distributions can only be analyzed asymptotically. This will not be discussed until Chapter **??**.

4.3.2 GLS Estimation

If the test for groupwise heteroskedasticity rejects the null hypothesis, one might believe that $\boldsymbol{\Sigma}_o$ is given by (4.9). Accordingly, the specified linear specification may be written as:

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \end{bmatrix},$$

where $\boldsymbol{y}_1$ is $T_1 \times 1$, $\boldsymbol{y}_2$ is $T_2 \times 1$, $\boldsymbol{X}_1$ is $T_1 \times k$, and $\boldsymbol{X}_2$ is $T_2 \times k$. A transformed specification is

$$\begin{bmatrix} \boldsymbol{y}_1/\sigma_1 \\ \boldsymbol{y}_2/\sigma_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1/\sigma_1 \\ \boldsymbol{X}_2/\sigma_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{e}_1/\sigma_1 \\ \boldsymbol{e}_2/\sigma_2 \end{bmatrix},$$

where the transformed y_t, $t = 1, \ldots, T$, have constant variance one. It follows that the GLS and FGLS estimators are, respectively,

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = \left[\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\sigma_1^2} + \frac{\boldsymbol{X}_2'\boldsymbol{X}_2}{\sigma_2^2}\right]^{-1} \left[\frac{\boldsymbol{X}_1'\boldsymbol{y}_1}{\sigma_1^2} + \frac{\boldsymbol{X}_2'\boldsymbol{y}_2}{\sigma_2^2}\right],$$

$$\hat{\boldsymbol{\beta}}_{\text{FGLS}} = \left[\frac{\boldsymbol{X}_1'\boldsymbol{X}_1}{\hat{\sigma}_1^2} + \frac{\boldsymbol{X}_2'\boldsymbol{X}_2}{\hat{\sigma}_2^2}\right]^{-1} \left[\frac{\boldsymbol{X}_1'\boldsymbol{y}_1}{\hat{\sigma}_1^2} + \frac{\boldsymbol{X}_2'\boldsymbol{y}_2}{\hat{\sigma}_2^2}\right],$$

where $\hat{\sigma}_{T_1}^2$ and $\hat{\sigma}_{T_2}^2$ are, again, the OLS variance estimates obtained from separate regressions using T_1 and T_2 observations, respectively. Observe that $\hat{\boldsymbol{\beta}}_{\text{FGLS}}$ is not a linear estimator in $\boldsymbol{y}$ so that its finite-sample properties are not clear.

If the Goldfeld-Quandt test rejects the null hypothesis, one might believe that $\sigma_t^2 = c\, x_{tj}^2$. A transformed specification is then

$$\frac{y_t}{x_{tj}} = \beta_j + \beta_1 \frac{1}{x_{tj}} + \cdots + \beta_{j-1}\frac{x_{t,j-1}}{x_{tj}} + \beta_{j+1}\frac{x_{t,j+1}}{x_{tj}} + \cdots + \beta_k \frac{x_{tk}}{x_{tj}} + \frac{e_t}{x_{tj}},$$

where $\text{var}(y_t/x_{tj}) = c := \sigma_o^2$. This is a very special case where the GLS estimator is readily computed as the OLS estimator for the transformed specification. Clearly, the validity of the GLS method crucially depends on whether the explanatory variable x_j can be correctly identified.

When $\sigma_t^2 = h(\alpha_0 + \boldsymbol{z}_t'\boldsymbol{\alpha}_1)$, it is typically difficult to implement an FGLS estimator, especially when h is nonlinear. If h is the identity function, one may regress the squared OLS residuals $\hat{e}_t^2$ on $\boldsymbol{z}_t$ to obtain estimates for α_0 and $\boldsymbol{\alpha}_1$. Of course, certain constraint must be imposed to ensure the fitted values are non-negative. The finite-sample properties of this estimator are difficult to analyze, however.

Remarks:

1. When a test for heteroskedasticity rejects the null hypothesis, there is really *no* guarantee that the alternative hypothesis (say, groupwise heteroskedasticity) must provide a correct description of $\mathrm{var}(y_t)$.

2. When a form of heteroskedasticity is incorrectly specified, it is likely that the resulting FGLS estimator is less efficient than the OLS estimator.

3. As discussed in Section 4.2.3, the finite-sample properties of FGLS estimators and hence the exact tests are usually not available. One may appeal to asymptotic theory to construct proper tests.

4.4 Serial Correlation

Another leading example that $\mathrm{var}(\boldsymbol{y}) \neq \sigma_o^2 \boldsymbol{I}_T$ is when the elements of $\boldsymbol{y}$ are correlated so that the off-diagonal elements of $\boldsymbol{\Sigma}_o$ are non-zero. This phenomenon is more common in time series data, though it is not necessary so. When time series data y_t are correlated over time, they are said to exhibit *serial correlation*. For cross-section data, the correlations of y_t are usually referred to as *spatial correlation*. We will concentrate on serial correlation.

4.4.1 A Simple Model of Serial Correlation

Consider time series y_t, $t = 1, \ldots, T$, with the constant variance σ_o^2. Then, the correlation coefficient between y_t and y_{t-i} is

$$\mathrm{corr}(y_t, y_{t-i}) = \frac{\mathrm{cov}(y_t, y_{t-i})}{\sqrt{\mathrm{var}(y_t)\,\mathrm{var}(y_{t-i})}} = \frac{\mathrm{cov}(y_t, y_{t-i})}{\sigma_o^2}, \qquad i = 0, 1, 2, \ldots, t-1;$$

in particular, $\mathrm{corr}(y_t, y_t) = 1$. Such correlations are also known as the *autocorrelations* of y_t. Similarly, $\mathrm{cov}(y_t, y_{t-i})$, $i = 0, 1, 2, \ldots, t-1$, are known as the *autocovariances* of y_t.

A very simple specification of autocovariances is

$$\mathrm{cov}(y_t, y_{t-i}) = \mathrm{cov}(y_t, y_{t+i}) = c^i \sigma_o^2,$$

where c is a constant, so that $\mathrm{corr}(y_t, y_{t-i}) = c^i$. That is, the autocovariances and autocorrelations depend only i, the time periods between two observations, but not on t. Moreover, the correlations between two observations decay exponentially fast when i increases. Equivalently, we may write

$$\mathrm{cov}(y_t, y_{t-i}) = c\ \mathrm{cov}(y_t, y_{t-i+1}).$$

Letting $\mathrm{corr}(y_t, y_{t-i}) = \rho_i$, we have

$$\rho_i = c\,\rho_{i-1}. \tag{4.10}$$

From this recursion we immediately see that $c = \rho_1$ which must be bounded between -1 and 1. It follows that $\mathrm{var}(\boldsymbol{y})$ is

$$\boldsymbol{\Sigma}_o = \sigma_o^2 \begin{bmatrix} 1 & \rho_1 & \rho_1^2 & \cdots & \rho_1^{T-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_1^{T-2} \\ \rho_1^2 & \rho_1 & 1 & \cdots & \rho_1^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_1^{T-1} & \rho_1^{T-2} & \rho_1^{T-3} & \cdots & 1 \end{bmatrix}. \tag{4.11}$$

To avoid singularity, ρ_1 cannot be ± 1.

A novel feature of this specification is that it, while permitting non-zero off-diagonal elements of $\boldsymbol{\Sigma}_o$, involves only two unknown parameters: σ_o^2 and ρ_1. The transformation matrix is then

$$\boldsymbol{\Sigma}_o^{-1/2} = \frac{1}{\sigma_o} \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -\frac{\rho_1}{\sqrt{1-\rho_1^2}} & \frac{1}{\sqrt{1-\rho_1^2}} & 0 & \cdots & 0 & 0 \\ 0 & -\frac{\rho_1}{\sqrt{1-\rho_1^2}} & \frac{1}{\sqrt{1-\rho_1^2}} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{\sqrt{1-\rho_1^2}} & 0 \\ 0 & 0 & 0 & \cdots & -\frac{\rho_1}{\sqrt{1-\rho_1^2}} & \frac{1}{\sqrt{1-\rho_1^2}} \end{bmatrix}.$$

Note that this choice of $\boldsymbol{\Sigma}_o^{-1/2}$ is not symmetric. As any matrix that is a constant proportion to $\boldsymbol{\Sigma}_o^{-1/2}$ can also serve as a transformation matrix for GLS estimation, the so-called *Cochrane-Orcutt Transformation* is based on

$$\boldsymbol{V}_o^{-1/2} = \sigma_o\sqrt{1-\rho_1^2}\,\boldsymbol{\Sigma}_o^{-1/2} = \begin{bmatrix} \sqrt{1-\rho_1^2} & 0 & 0 & \cdots & 0 & 0 \\ -\rho_1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho_1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -\rho_1 & 1 \end{bmatrix},$$

which depends only on ρ_1.

The data from the Cochrane-Orcutt transformation are

$$\begin{aligned} y_1^* &= (1-\rho_1^2)^{1/2} y_1, \qquad & \boldsymbol{x}_1^* &= (1-\rho_1^2)^{1/2}\boldsymbol{x}_1, \\ y_t^* &= y_t - \rho_1 y_{t-1}, & \boldsymbol{x}_t^* &= \boldsymbol{x}_t - \rho_1 \boldsymbol{x}_{t-1}, \qquad t = 2, \cdots, T, \end{aligned}$$

where $\boldsymbol{x}_t$ is the t thcolumn of $\boldsymbol{X}'$. It is then clear that

$$\begin{aligned} &\operatorname{var}(y_1^*) = (1-\rho_1^2)\sigma_o^2, \\ &\operatorname{var}(y_t^*) = \sigma_o^2 + \rho_1^2\sigma_o^2 - 2\rho_1^2\sigma_o^2 = (1-\rho_1^2)\sigma_o^2, \qquad t = 2, \ldots, T. \end{aligned}$$

Moreover, for each i,

$$\begin{aligned} \operatorname{cov}(y_t^*, y_{t-i}^*) &= \operatorname{cov}(y_t, y_{t-i}) - \rho_1 \operatorname{cov}(y_{t-1}, y_{t-i}) - \rho_1 \operatorname{cov}(y_t, y_{t-i-1}) \\ &\qquad - \rho_1^2 \operatorname{cov}(y_{t-1}, y_{t-i-1}) \\ &= 0. \end{aligned}$$

Hence, the transformed variable y_t^* satisfies the classical conditions, as it ought to be. Then provided that ρ_1 is known, regressing y_t^* on $\boldsymbol{x}_t^*$ yields the GLS estimator for $\boldsymbol{\beta}_o$.

4.4.2 An Alternative View

There is an alternative approach to generate the variance-covariance matrix (4.11). Under [A2](i), let

$$\boldsymbol{\epsilon} := \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o.$$

The vector $\boldsymbol{\epsilon}$ is usually referred to as the vector of *disturbances*. Note that $\boldsymbol{\epsilon}$ is not the same as the residual vector $\hat{\boldsymbol{e}}$. While the former is not observable because $\boldsymbol{\beta}_o$ is unknown, the later is obtained from OLS estimation and hence observable. Under [A2], $\mathbb{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and

$$\operatorname{var}(\boldsymbol{y}) = \operatorname{var}(\boldsymbol{\epsilon}) = \mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}').$$

The variance and covariance structure of $\boldsymbol{y}$ is thus the same as that of $\boldsymbol{\epsilon}$.

A time series is said to be *weakly stationary* if its mean, variance, and autocovariances are all independent of the time index t. Thus, a weakly stationary series cannot exhibit trending behavior and has relatively stable fluctuations. In particular, a time series with zero mean, a constant variance, and zero autocovariances is weakly stationary and also known as a *white noise*. Let $\{u_t\}$ be a white noise with $\mathbb{E}(u_t) = 0$, $\mathbb{E}(u_t^2) = \sigma_u^2$,

and $\mathbb{E}(u_t u_\tau) = 0$ for $t \neq \tau$. Now suppose that the elements of $\boldsymbol{\epsilon}$ is generated as a weakly stationary AR(1) process (*autoregressive process of order 1*):

$$\epsilon_t = \alpha_1 \epsilon_{t-1} + u_t, \tag{4.12}$$

with $\epsilon_0 = 0$. By recursive substitution, (4.12) can be expressed as

$$\epsilon_t = \sum_{i=0}^{t-1} \alpha_1^i u_{t-i}, \tag{4.13}$$

a weighted sum of current and previous random innovations (shocks).

It follows from (4.13) that $\mathbb{E}(\epsilon_t) = 0$ and $\mathbb{E}(u_t, \epsilon_{t-s}) = 0$ for all t and $s \geq 1$. By weak stationarity, $\operatorname{var}(\epsilon_t)$ is a constant, so that for all t,

$$\operatorname{var}(\epsilon_t) = \alpha_1^2 \operatorname{var}(\epsilon_{t-1}) + \sigma_u^2 = \sigma_u^2/(1 - \alpha_1^2).$$

Clearly, the right-hand side would not be meaningful unless $|\alpha_1| < 1$. The autocovariance of ϵ_t and ϵ_{t-1} is, by weak stationarity,

$$\mathbb{E}(\epsilon_t \epsilon_{t-1}) = \alpha_1 \mathbb{E}(\epsilon_{t-1}^2) = \alpha_1 \frac{\sigma_u^2}{1 - \alpha_1^2}.$$

This shows that

$$\alpha_1 = \operatorname{corr}(\epsilon_t, \epsilon_{t-1}) = \operatorname{corr}(y_t, y_{t-1}) = \rho_1.$$

Similarly,

$$\mathbb{E}(\epsilon_t \epsilon_{t-2}) = \alpha_1 \mathbb{E}(\epsilon_{t-1} \epsilon_{t-2}) = \alpha_1^2 \frac{\sigma_u^2}{1 - \alpha_1^2},$$

so that

$$\operatorname{corr}(\epsilon_t, \epsilon_{t-2}) = \alpha_1 \operatorname{corr}(\epsilon_t, \epsilon_{t-1}) = \rho_1^2.$$

More generally, we can write for $i = 1, 2, \ldots,$

$$\operatorname{corr}(\epsilon_t, \epsilon_{t-i}) = \rho_1 \operatorname{corr}(\epsilon_t, \epsilon_{t-i+1}) = \rho_1^i,$$

which depend only on i, the time difference between two ϵ's, but not on t. This is precisely what we postulated in (4.10). The variance-covariance matrix $\boldsymbol{\Sigma}_o$ under this structure is also (4.11) with $\sigma_o^2 = \sigma_u^2/(1 - \rho_1^2)$.

The AR(1) structure of disturbances also permits a straightforward extension. Consider the disturbances that are generated as an AR(p) process (autoregressive process of order p):

$$\epsilon_t = \alpha_1 \epsilon_{t-1} + \cdots + \alpha_p \epsilon_{t-p} + u_t, \tag{4.14}$$

where the coefficients $\alpha_1, \ldots, \alpha_p$ should also be restricted to ensure weak stationarity; we omit the details. Of course, ϵ_t may follow different structures and are still serially correlated. For example, ϵ_t may be generated as an MA(1) process (*moving average process of order 1*):

$$\epsilon_t = u_t + \alpha_1 u_{t-1}, \qquad |\alpha_1| < 1,$$

where $\{u_t\}$ is a white noise; see e.g., Exercise 4.5.

4.4.3 Tests for AR(1) Disturbances

As the AR(1) structure of disturbances is one of the most commonly used specification of serial correlation, we now consider the tests of the null hypothesis of no serial correlation ($\alpha_1 = \rho_1 = 0$) against AR(1) disturbances. We discuss only the celebrated Durbin-Watson test and Durbin's h test; the discussion of other large-sample tests will be deferred to Chapter 6.

In view of the AR(1) structure, a natural estimator of ρ_1 is the OLS estimator of regressing the OLS residual $\hat{e}_t$ on its immediate lag $\hat{e}_{t-1}$:

$$\hat{\rho}_T = \frac{\sum_{t=2}^T \hat{e}_t \hat{e}_{t-1}}{\sum_{t=2}^T \hat{e}_{t-1}^2}. \tag{4.15}$$

The Durbin-Watson statistic is

$$d = \frac{\sum_{t=2}^T (\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^T \hat{e}_t^2}.$$

When the sample size T is large, it can be seen that

$$\begin{aligned} d &= 2 - 2\hat{\rho}_T \frac{\sum_{t=2}^T \hat{e}_{t-1}^2}{\sum_{t=1}^T \hat{e}_t^2} - \frac{\hat{e}_1^2 + \hat{e}_T^2}{\sum_{t=1}^T \hat{e}_t^2} \\ &\approx 2(1 - \hat{\rho}_T). \end{aligned}$$

For $0 < \hat{\rho}_T \leq 1$ ($-1 \leq \hat{\rho}_T < 0$), the Durbin-Watson statistic is such that $0 \leq d < 2$ ($2 < d \leq 4$), which suggests that there is some positive (negative) serial correlation.

Hence, this test essentially checks whether $\hat{\rho}_T$ is sufficiently "close" to zero (i.e., d is close to 2).

A major difficulty of the Durbin-Watson test is that the exact null distribution of d depends on the matrix $\boldsymbol{X}$ and therefore varies with data. As such, the critical values of d cannot be tabulated. Nevertheless, it has been shown that the null distribution of d lies between the distributions of a lower bound (d_L) and an upper bound (d_U) in the following sense. Given the significance level α, let d^*_α, $d^*_{L,\alpha}$ and $d^*_{U,\alpha}$ denote, respectively, the critical values of d, d_L and d_U. For example, $\mathbb{P}\{d < d^*_\alpha\} = \alpha$). Then for each α, $d^*_{L,\alpha} < d^*_\alpha < d^*_{U,\alpha}$. While the distribution of d is data dependent, the distributions of d_L and d_U are independent of $\boldsymbol{X}$. Thus, the critical values $d^*_{L,\alpha}$ and $d^*_{U,\alpha}$ can be tabulated. One may rely on these critical values to construct a "conservative" decision rule.

Specifically, when the alternative hypothesis is $\rho_1 > 0$ ($\rho_1 < 0$), the decision rule of the Durbin-Watson test is:

(1) Reject the null if $d < d^*_{L,\alpha}$ ($d > 4 - d^*_{L,\alpha}$).

(2) Do not reject the null if $d > d^*_{U,\alpha}$ ($d < 4 - d^*_{U,\alpha}$).

(3) Test is inconclusive if $d^*_{L,\alpha} < d < d^*_{U,\alpha}$ ($4 - d^*_{L,\alpha} > d > 4 - d^*_{U,\alpha}$).

This is not completely satisfactory because the test may yield no conclusion. Some econometric packages such as SHAZAM now compute the exact Durbin-Watson distribution for each regression and report the exact p-values. When such as program is available, this test does not have to rely on the critical values of d_L and d_U, and it is always conclusive. Note that the tabulated critical values of the Durbin-Watson statistic are for the specifications with a constant term; the critical values for the specifications without a constant term can be found in Farebrother (1980).

Another problem with the Durbin-Watson statistic is that its null distribution holds only under the classical conditions [A1] and [A3]. In the time series context, it is quite common to include a lagged dependent variable as a regressor so that [A1] is violated. A leading example is the specification

$$y_t = \beta_1 + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \gamma y_{t-1} + e_t.$$

This model can also be derived from certain behavioral assumptions; see Exercise 4.6. It has been shown that the Durbin-Watson statistic under this specification is biased toward 2. That is, this test would not reject the null hypothesis even when serial correlation is present. On the other hand, Durbin's h test is designed specifically for

the specifications that contain a lagged dependent variable. Let $\hat{\gamma}_T$ be the OLS estimate of γ and $\widehat{\text{var}}(\hat{\gamma}_T)$ be the OLS estimate of $\text{var}(\hat{\gamma}_T)$. The h statistic is

$$h = \hat{\rho}_T \sqrt{\frac{T}{1 - T\widehat{\text{var}}(\hat{\gamma}_T)}},$$

and its asymptotic null distribution is $N(0,1)$. A clear disadvantage of Durbin's h test is that it cannot be calculated when $\widehat{\text{var}}(\hat{\gamma}_T) \geq 1/T$. This test can also be derived as a Lagrange Multiplier test; see Chapter **??**

If we have quarterly data and want to test for the fourth-order serial correlation, the statistic analogous to the Durbin-Watson statistic is

$$d_4 = \frac{\sum_{t=5}^{T}(\hat{e}_t - \hat{e}_{t-4})^2}{\sum_{t=1}^{T}\hat{e}_t^2};$$

see Wallis (1972) for corresponding critical values.

4.4.4 FGLS Estimation

Recall that $\mathbf{\Sigma}_o$ depends on two parameters σ_o^2 and ρ_1. We may use a generic notation $\mathbf{\Sigma}(\sigma^2, \rho)$ to denote this function of σ^2 and ρ. In particular, $\mathbf{\Sigma}_o = \mathbf{\Sigma}(\sigma_o^2, \rho_1)$. Similarly, we may also write $\boldsymbol{V}(\rho)$ such that $\boldsymbol{V}_o = \boldsymbol{V}(\rho_1)$. The transformed data based on $\boldsymbol{V}(\rho)^{-1/2}$ are

$$\begin{aligned} y_1(\rho) &= (1-\rho^2)^{1/2} y_1, \qquad & \boldsymbol{x}_1(\rho) &= (1-\rho^2)^{1/2}\boldsymbol{x}_1, \\ y_t(\rho) &= y_t - \rho y_{t-1}, & \boldsymbol{x}_t(\rho) &= \boldsymbol{x}_t - \rho\boldsymbol{x}_{t-1}, \qquad t = 2, \cdots, T. \end{aligned}$$

Hence, $y_t^* = y_t(\rho_1)$ and $\boldsymbol{x}_t^* = \boldsymbol{x}_t(\rho_1)$.

To obtain an FGLS estimator, we must first estimate ρ_1 by some estimator $\hat{\rho}_T$ and then construct the transformation matrix as $\hat{\boldsymbol{V}}_T^{-1/2} = \boldsymbol{V}(\hat{\rho}_T)^{-1/2}$. Here, $\hat{\rho}_T$ may be computed as in (4.15); other estimators for ρ_1 may also be used, e.g., $\check{\rho}_T = \hat{\rho}_T(T-k)/(T-1)$. The transformed data are then $y_t(\hat{\rho}_T)$ and $\boldsymbol{x}_t(\hat{\rho}_T)$. An FGLS estimator is obtained by regressing $y_t(\hat{\rho}_T)$ on $\boldsymbol{x}_t(\hat{\rho}_T)$. Such an estimator is known as the Prais-Winsten estimator or the Cochrane-Orcutt estimator when the first observation is dropped in computation.

The following iterative procedure is also commonly employed in practice.

(1) Perform OLS estimation and compute $\hat{\rho}_T$ as in (4.15) using the OLS residuals $\hat{e}_t$.

(2) Perform the Cochrane-Orcutt transformation based on $\hat{\rho}_T$ and compute the resulting FGLS estimate $\hat{\boldsymbol{\beta}}_{\text{FGLS}}$ by regressing $y_t(\hat{\rho}_T)$ on $\boldsymbol{x}_t(\hat{\rho}_T)$.

(3) Compute a new $\hat{\rho}_T$ as in (4.15) with $\hat{e}_t$ replaced by the FGLS residuals

$$\hat{e}_{t,\text{FGLS}} = y_t - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_{\text{FGLS}}.$$

(4) Repeat steps (2) and (3) until $\hat{\rho}_T$ converges numerically, i.e., when $\hat{\rho}_T$ from two consecutive iterations differ by a value smaller than a pre-determined convergence criterion.

Note that steps (1) and (2) above already generate an FGLS estimator. More iterations do not improve the asymptotic properties of the resulting estimator but may have a significant effect in finite samples. This procedure can be extended easily to estimate the specification with higher-order AR disturbances.

Alternatively, the Hildreth-Lu procedure adopts *grid search* to find the $\rho_1 \in (-1, 1)$ that minimizes the sum of squared errors of the model. This procedure is computationally intensive, and it is difficult to implement when ϵ_t have an AR(p) structure with $p > 2$.

In view of the log-likelihood function (4.7), we must compute $\det(\boldsymbol{\Sigma}_o)$. Clearly,

$$\det(\boldsymbol{\Sigma}_o) = \frac{1}{\det(\boldsymbol{\Sigma}_o^{-1})} = \frac{1}{[\det(\boldsymbol{\Sigma}_o^{-1/2})]^2}.$$

In terms of the notations in the AR(1) formulation, $\sigma_o^2 = \sigma_u^2/(1-\rho_1^2)$, and

$$\boldsymbol{\Sigma}_o^{-1/2} = \frac{1}{\sigma_o\sqrt{1-\rho_1^2}}\boldsymbol{V}_o^{-1/2} = \frac{1}{\sigma_u}\boldsymbol{V}_o^{-1/2}.$$

As $\det(\boldsymbol{V}_o^{-1/2}) = (1-\rho_1^2)^{1/2}$, we then have

$$\det(\boldsymbol{\Sigma}_o) = (\sigma_u^2)^T(1-\rho_1^2)^{-1}.$$

The log-likelihood function for given σ_u^2 and ρ_1 is

$$\begin{aligned}&\log L(\boldsymbol{\beta}; \sigma_u^2, \rho_1)\\ &\quad = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma_u^2) + \frac{1}{2}\log(1-\rho_1^2) - \frac{1}{2\sigma_u^2}(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta})'(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}).\end{aligned}$$

Clearly, when σ_u^2 and ρ_1 are known, the MLE of $\boldsymbol{\beta}$ is just the GLS estimator.

If σ_u^2 and ρ_1 are unknown, the log-likelihood function reads:

$$\begin{aligned}&\log L(\boldsymbol{\beta}, \sigma^2, \rho)\\ &\quad = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) + \frac{1}{2}\log(1-\rho^2) - \frac{1}{2\sigma^2}(1-\rho^2)(y_1 - \boldsymbol{x}_1'\boldsymbol{\beta})^2\\ &\qquad - \frac{1}{2\sigma^2}\sum_{t=2}^{T}[(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}) - \rho(y_{t-1} - \boldsymbol{x}_{t-1}'\boldsymbol{\beta})]^2,\end{aligned}$$

which is a nonlinear function of the parameters. Nonlinear optimization methods are therefore needed to compute the MLEs of $\boldsymbol{\beta}$, σ^2, and ρ. For a given $\boldsymbol{\beta}$, estimating ρ by regressing $e_t(\boldsymbol{\beta}) = y_t - \boldsymbol{x}_t'\boldsymbol{\beta}$ on $e_{t-1}(\boldsymbol{\beta})$ is equivalent to maximizing the last term of the log-likelihood function above. This does not yield an MLE because the other terms involving ρ, namely,

$$\frac{1}{2}\log(1-\rho^2) - \frac{1}{2\sigma^2}(1-\rho^2)(y_1 - \boldsymbol{x}_1'\boldsymbol{\beta})^2,$$

have been ignored. This shows that the aforementioned iterative procedure does not result in the MLEs.

Remark: Exact tests based on FGLS estimation results are not available because the finite-sample distribution of the FGLS estimator is, again, unknown. Asymptotic theory is needed to construct proper tests.

4.5 Linear Probability Model

In some applications researchers are interested in analyzing why consumers own a house or participate a particular event. The ownership or the choice of participation are typically represented by a *binary* variable that takes the values one and zero. If the dependent variable in a linear regression is binary, we will see below that both the OLS and FGLS methods are not appropriate.

Let $\boldsymbol{x}_t$ denote the t^{th} column of $\boldsymbol{X}'$. The t^{th} observation of the linear specification $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ can be expressed as

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t.$$

For the binary dependent variable y whose t^{th} observation is $y_t = 1$ or 0, we know

$$\mathbb{E}(y_t) = \mathbb{P}(y_t = 1).$$

Thus, $\boldsymbol{x}_t'\boldsymbol{\beta}$ is just a specification of the probability that $y_t = 1$. As such, the linear specification of binary dependent variables is usually referred to as the *linear probability model*.

When [A1] and [A2](i) hold for a linear probability model,

$$\mathbb{E}(y_t) = \mathbb{P}(y_t = 1) = \boldsymbol{x}_t'\boldsymbol{\beta}_o,$$

and the OLS estimator is unbiased for $\boldsymbol{\beta}_o$. Note, however, that the variance of y_t is

$$\text{var}(y_t) = \mathbb{P}(y_t = 1)[1 - \mathbb{P}(y_t = 1)].$$

Under [A1] and [A2](i),

$$\text{var}(y_t) = \boldsymbol{x}_t'\boldsymbol{\beta}_o(1 - \boldsymbol{x}_t'\boldsymbol{\beta}_o),$$

which varies with $\boldsymbol{x}_t$. Thus, the linear probability model suffers from the problem of heteroskedasticity, and the OLS estimator is not the BLUE for $\boldsymbol{\beta}_o$. Apart from the efficiency issue, the OLS method is still not appropriate for the linear probability model because the OLS fitted values need not be bounded between zero and one. When $\boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_T$ is negative or greater than one, it cannot be interpreted as a probability and hence becomes meaningless.

Although the GLS estimator is the BLUE, it is not available because $\boldsymbol{\beta}_o$, and hence $\text{var}(y_t)$, is unknown. Nevertheless, if y_t are uncorrelated so that $\text{var}(\boldsymbol{y})$ is diagonal, an FGLS estimator may be obtained using the transformation matrix

$$\hat{\boldsymbol{\Sigma}}_T^{-1/2} = \text{diag}\left[\boldsymbol{x}_1'\hat{\boldsymbol{\beta}}_T(1 - \boldsymbol{x}_1'\hat{\boldsymbol{\beta}}_T)]^{-1/2}, \boldsymbol{x}_2'\hat{\boldsymbol{\beta}}_T(1 - \boldsymbol{x}_2'\hat{\boldsymbol{\beta}}_T)]^{-1/2}, \ldots, \right.$$
$$\left. \boldsymbol{x}_T'\hat{\boldsymbol{\beta}}_T(1 - \boldsymbol{x}_T'\hat{\boldsymbol{\beta}}_T)]^{-1/2}\right],$$

where $\hat{\boldsymbol{\beta}}_T$ is the OLS estimator of $\boldsymbol{\beta}_o$. Such an estimator breaks down when $\hat{\boldsymbol{\Sigma}}_T^{-1/2}$ is not available (i.e., when $\boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_T$ is negative or greater than one). Even when $\hat{\boldsymbol{\Sigma}}_T^{-1/2}$ can be computed, there is still no guarantee that the FGLS fitted values are bounded between zero and one. This shows that the FGLS method may not always be a solution when the OLS method fails.

This example also illustrates the importance of data characteristics in estimation and modeling. Without taking into account the binary nature of the dependent variable, even the FGLS method may be invalid. More appropriate methods for specifications with binary dependent variables will be discussed in Chapter **??**.

4.6 Seemingly Unrelated Regressions

In many econometric practices, it is also important to jointly study the behavior of several dependent variables. For example, the input demands of an firm may be described using a system of linear regression functions in which each regression represents the demand function of a particular input.

Consider the specification of a system of N equations, each with k_i explanatory variables and T observations. Specifically,

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{e}_i, \qquad i = 1, 2, \ldots, N, \tag{4.16}$$

where for each i, $\boldsymbol{y}_i$ is $T \times 1$, $\boldsymbol{X}_i$ is $T \times k_i$, and $\boldsymbol{\beta}_i$ is $k_i \times 1$. The system (4.16) is also known as a specification of *seemingly unrelated regressions* (SUR). Stacking the equations of (4.16) yields

$$\underbrace{\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_N \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{X}_N \end{bmatrix}}_{\boldsymbol{X}} \underbrace{\begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_N \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_N \end{bmatrix}}_{\boldsymbol{e}}. \tag{4.17}$$

This is a linear specification (3.1) with $k = \sum_{i=1}^{N} k_i$ explanatory variables and TN observations. It is not too hard to see that the whole system (4.17) satisfies the identification requirement whenever every specification of (4.16) does.

Suppose that the classical conditions [A1] and [A2] hold for each specified linear regression in the system. Then under [A2](i), there exists $\boldsymbol{\beta}_o = (\boldsymbol{\beta}_{o,1}' \ \ldots \ \boldsymbol{\beta}_{o,N}')'$ such that $\mathbb{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}_o$. The OLS estimator obtained from (4.17) is therefore unbiased. Note, however, that [A2](ii) for each linear regression ensures only that, for each i,

$$\operatorname{var}(\boldsymbol{y}_i) = \sigma_i^2 \boldsymbol{I}_T;$$

there is no restriction on the correlations between $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$. The variance-covariance matrix of $\boldsymbol{y}$ is then

$$\operatorname{var}(\boldsymbol{y}) = \boldsymbol{\Sigma}_o = \begin{bmatrix} \sigma_1^2 \boldsymbol{I}_T & \operatorname{cov}(\boldsymbol{y}_1, \boldsymbol{y}_2) & \cdots & \operatorname{cov}(\boldsymbol{y}_1, \boldsymbol{y}_N) \\ \operatorname{cov}(\boldsymbol{y}_2, \boldsymbol{y}_1) & \sigma_2^2 \boldsymbol{I}_T & \cdots & \operatorname{cov}(\boldsymbol{y}_2, \boldsymbol{y}_N) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(\boldsymbol{y}_N, \boldsymbol{y}_1) & \operatorname{cov}(\boldsymbol{y}_N, \boldsymbol{y}_2) & \cdots & \sigma_N^2 \boldsymbol{I}_T \end{bmatrix}. \tag{4.18}$$

That is, the vector of stacked dependent variables violates [A2](ii), even when each individual dependent variable has a scalar variance-covariance matrix. Consequently, the OLS estimator of the whole system, $\hat{\boldsymbol{\beta}}_{TN} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$, is not the BLUE in general. In fact, owing to the block-diagonal structure of $\boldsymbol{X}$, $\hat{\boldsymbol{\beta}}_{TN}$ simply consists of N equation-by-equation OLS estimators and hence ignores the correlations between equations and heteroskedasticity across equations.

In practice, it is also typical to postulate that for $i \neq j$,

$$\operatorname{cov}(\boldsymbol{y}_i, \boldsymbol{y}_j) = \sigma_{ij} \boldsymbol{I}_T,$$

that is, y_{it} and y_{jt} are contemporaneously correlated but y_{it} and $y_{j\tau}$, $t \neq \tau$, are serially uncorrelated. Under this condition, (4.18) simplifies to $\boldsymbol{\Sigma}_o = \boldsymbol{S}_o \otimes \boldsymbol{I}_T$ with

$$\boldsymbol{S}_o = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1N} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \cdots & \sigma_N^2 \end{bmatrix}.$$

As $\boldsymbol{\Sigma}_o^{-1} = \boldsymbol{S}_o^{-1} \otimes \boldsymbol{I}_T$, the GLS estimator of (4.17) is

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = [\boldsymbol{X}'(\boldsymbol{S}_o^{-1} \otimes \boldsymbol{I}_T)\boldsymbol{X}]^{-1}\boldsymbol{X}'(\boldsymbol{S}_o^{-1} \otimes \boldsymbol{I}_T)\boldsymbol{y},$$

and its covariance matrix is $[\boldsymbol{X}'(\boldsymbol{S}_o^{-1} \otimes \boldsymbol{I}_T)\boldsymbol{X}]^{-1}$.

It is readily verified that when $\sigma_{ij} = 0$ for all $i \neq j$, $\boldsymbol{S}_o$ becomes a diagonal matrix, and so is $\boldsymbol{\Sigma}_o$. The resulting GLS estimator thus reduces to the OLS estimator. This should not be too surprising because estimating the SUR system would be unnecessary if the dependent variables are in fact uncorrelated. (Note that the heteroskedasticity across equations does not affect this result.) If all the equations in the system have the same regressors, i.e., $\boldsymbol{X}_i = \boldsymbol{X}_0$ (say), the GLS estimator is also the same as the OLS estimator; see e.g., Exercise 4.7. More generally, it can be shown that there would not be much efficiency gain for GLS estimation if $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ are less correlated and/or $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ are highly correlated; see e.g., Goldberger (1991, p. 328) for an illustrative example.

The FGLS estimator can be obtained by replacing $\boldsymbol{S}_o^{-1}$ with $\hat{\boldsymbol{S}}_{TN}^{-1}$, where $\hat{\boldsymbol{S}}_{TN}$ is an $N \times N$ matrix computed as

$$\hat{\boldsymbol{S}}_{TN} = \frac{1}{T} \begin{bmatrix} \hat{\boldsymbol{e}}_1' \\ \hat{\boldsymbol{e}}_2' \\ \vdots \\ \hat{\boldsymbol{e}}_N' \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{e}}_1 & \hat{\boldsymbol{e}}_2 & \cdots & \hat{\boldsymbol{e}}_N \end{bmatrix},$$

where $\hat{\boldsymbol{e}}_i$ is the OLS residual vector of the i^{th} equation. The elements of this matrix are

$$\hat{\sigma}_i^2 = \frac{\hat{\boldsymbol{e}}_i'\hat{\boldsymbol{e}}_i}{T}, \qquad i = 1, \ldots, N,$$

$$\hat{\sigma}_{ij} = \frac{\hat{\boldsymbol{e}}_i'\hat{\boldsymbol{e}}_j}{T}, \qquad i \neq j,\ i, j = 1, \ldots, N.$$

Note that $\hat{\boldsymbol{S}}_{TN}$ is of an outer product form and hence a positive semi-definite matrix. One may also replace the denominator of $\hat{\sigma}_i^2$ with $T - k_i$ and the denominator of $\hat{\sigma}_{ij}$

with $T - \max(k_i, k_j)$. The resulting estimator $\hat{\boldsymbol{S}}_{TN}$ need not be positive semi-definite, however.

Remark: The estimator $\hat{\boldsymbol{S}}_{TN}$ mentioned above is valid provided that $\text{var}(\boldsymbol{y}_i) = \sigma_i^2 \boldsymbol{I}_T$ and $\text{cov}(\boldsymbol{y}_i, \boldsymbol{y}_j) = \sigma_{ij} \boldsymbol{I}_T$. If these assumptions do not hold, FGLS estimation would be much more complicated. This may happen when heteroskedasticity and serial correlations are present in each equation, or when $\text{cov}(y_{it}, y_{jt})$ changes over time.

4.7 Models for Panel Data

A data set that contains N cross-section units (individuals, families, firms, or countries), each with some time-series observations, is known as a *panel data* set. Well known panel data sets in the U.S. include the National Longitudinal Survey (NLS) of Labor Market Experience and the Michigan Panel Study of Income Dynamics (PSID). Building these data sets is very costly because they are obtained by tracking thousands of individuals through time. Some panel data may be easier to establish; for example, the GDP data for all G7 countries over 30 years also form a panel data set. Panel data permit analysis of topics that could not be studied using only cross-section or time-series data. In this section, we are mainly concerned with the panel data set that involves a large number of cross-section units, each with a short time series.

4.7.1 Fixed Effects Model

Given a panel data set, the basic linear specification allowing for individual effects (i.e., effects that are changing across individual units but remain constant over time) is

$$y_{it} = \boldsymbol{x}_{it}' \boldsymbol{\beta}_i + e_{it}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T,$$

where $\boldsymbol{x}_{it}$ is $k \times 1$ and $\boldsymbol{\beta}_i$ depends only on i but not on t. Clearly, there is no time-specific effect in this specification; this may be reasonable when only a short time series is observed for each individual unit.

Analogous to the notations in the SUR system (4.16), we can also write the specification above as

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta}_i + \boldsymbol{e}_i, \qquad i = 1, 2, \ldots, N, \tag{4.19}$$

where $\boldsymbol{y}_i$ is $T \times 1$, $\boldsymbol{X}_i$ is $T \times k$, and $\boldsymbol{e}_i$ is $T \times 1$. This is again a complex system involving $k \times N$ parameters. Here, the dependent variable y and explanatory variables X are the same across individual units such that $\boldsymbol{y}_i$ and $\boldsymbol{X}_i$ are simply their observations for each

individual i. For example, y may be the family consumption expenditure, and each y_i contains family i's annual consumption expenditures. By contrast, $\boldsymbol{y}_i$ and $\boldsymbol{X}_i$ may be different variables in a SUR system.

When T is small (i.e., observed time series are short), estimating (4.19) is not feasible. A simpler form of (4.19) is such that only the intercept terms change with i and the other parameters remain constant across i:

$$\boldsymbol{y}_i = \boldsymbol{\ell}_T a_i + \boldsymbol{Z}_i \boldsymbol{b} + \boldsymbol{e}_i, \qquad i = 1, 2, \ldots, N, \tag{4.20}$$

where $\boldsymbol{\ell}_T$ is the T-dimensional vector of ones, $[\boldsymbol{\ell}_T \ \boldsymbol{Z}_i] = \boldsymbol{X}_i$ and $[a_i \ \boldsymbol{b}']' = \boldsymbol{\beta}_i$. Thus, individual effects are completely captured by the intercept terms in (4.20). This simplifies (4.19) from kN to $N + k - 1$ parameters. Note that this specification treats a_i as non-random parameters and is known as the *fixed effects model*. Stacking N equations in (4.20) together we obtain

$$\underbrace{\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_N \end{bmatrix}}_{\boldsymbol{y}} = \underbrace{\begin{bmatrix} \boldsymbol{\ell}_T & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\ell}_T & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{\ell}_T \end{bmatrix}}_{\boldsymbol{D}} \underbrace{\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix}}_{\boldsymbol{a}} + \underbrace{\begin{bmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \\ \vdots \\ \boldsymbol{Z}_N \end{bmatrix}}_{\boldsymbol{Z}} \boldsymbol{b} + \underbrace{\begin{bmatrix} \boldsymbol{e}_1 \\ \boldsymbol{e}_2 \\ \vdots \\ \boldsymbol{e}_N \end{bmatrix}}_{\boldsymbol{e}}. \tag{4.21}$$

This is just a linear specification (3.1) with $N + k - 1$ explanatory variables and TN observations. Note that each column of $\boldsymbol{D}$ is in effect a dummy variable for the i^{th} individual unit. In what follows, an individual unit will be referred to as a "group."

Let $\boldsymbol{z}_{it}$ denote the t^{th} column of $\boldsymbol{Z}_i'$, where $\boldsymbol{Z}_i'$ is the i^{th} block of $\boldsymbol{Z}'$. For $\boldsymbol{z}_{it}$, the i^{th} group average over time is

$$\bar{\boldsymbol{z}}_i = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{z}_{it} = \frac{1}{T} \boldsymbol{Z}_i' \boldsymbol{\ell}_T;$$

for y_{it}, the group average over time is

$$\bar{\boldsymbol{y}}_i = \frac{1}{T} \boldsymbol{y}_i' \boldsymbol{\ell}_T.$$

The overall sample average of $\boldsymbol{z}_{it}$ (average over time and groups) is

$$\bar{\boldsymbol{z}} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \boldsymbol{z}_{it} = \frac{1}{NT} \boldsymbol{Z}' \boldsymbol{\ell}_{NT},$$

and the overall sample average of y_{it} is

$$\bar{\boldsymbol{y}} = \frac{1}{NT} \boldsymbol{y}' \boldsymbol{\ell}_{NT}.$$

Note that

$$\bar{\boldsymbol{z}} = \frac{1}{N}\sum_{i=1}^{N} \bar{\boldsymbol{z}}_i,$$

so that the overall sample average is simply the sample mean of all group averages. Similarly, $\bar{\boldsymbol{y}}$ is also the sample average of $\bar{\boldsymbol{y}}_i$.

It can be seen that the specification (4.21) satisfies the identification requirement provided that no column of $\boldsymbol{Z}_i$ is a constant (i.e., there is no time invariant regressor for each group). Once the identification requirement is satisfied, the OLS estimator can be computed. By Theorem 3.3, the OLS estimator for $\boldsymbol{b}$ is

$$\hat{\boldsymbol{b}}_{NT} = [\boldsymbol{Z}'(\boldsymbol{I}_{NT} - \boldsymbol{P}_D)\boldsymbol{Z}]^{-1}\boldsymbol{Z}'(\boldsymbol{I}_{NT} - \boldsymbol{P}_D)\boldsymbol{y}, \tag{4.22}$$

where $\boldsymbol{P}_D = \boldsymbol{D}(\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'$ is a projection matrix. Thus, $\hat{b}_{NT}$ can be obtained by regressing $(\boldsymbol{I}_{NT} - \boldsymbol{P}_D)\boldsymbol{y}$ on $(\boldsymbol{I}_{NT} - \boldsymbol{P}_D)\boldsymbol{Z}$.

More specifically, write $\boldsymbol{D} = \boldsymbol{I}_N \otimes \boldsymbol{\ell}_T$, so that

$$\boldsymbol{P}_D = (\boldsymbol{I}_N \otimes \boldsymbol{\ell}_T)(\boldsymbol{I}_N \otimes \boldsymbol{\ell}_T'\boldsymbol{\ell}_T)^{-1}(\boldsymbol{I}_N \otimes \boldsymbol{\ell}_T') = \boldsymbol{I}_N \otimes \boldsymbol{\ell}_T\boldsymbol{\ell}_T'/T.$$

It follows that $\boldsymbol{I}_{NT} - \boldsymbol{P}_D = \boldsymbol{I}_N \otimes (\boldsymbol{I}_T - \boldsymbol{\ell}_T\boldsymbol{\ell}_T'/T)$ and that

$$(\boldsymbol{I}_{NT} - \boldsymbol{P}_D)\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_N \end{pmatrix} - \begin{pmatrix} \boldsymbol{\ell}_T\bar{\boldsymbol{y}}_1 \\ \boldsymbol{\ell}_T\bar{\boldsymbol{y}}_2 \\ \vdots \\ \boldsymbol{\ell}_T\bar{\boldsymbol{y}}_N \end{pmatrix},$$

where $\bar{\boldsymbol{y}}_i$ is the group average of the elements in $\boldsymbol{y}_i$. The t^{th} observation in the i^{th} block of $(\boldsymbol{I}_{NT} - \boldsymbol{P}_D)\boldsymbol{y}$ is then $y_{it} - \bar{\boldsymbol{y}}_i$, the deviation of y_{it} from its group average. Similarly,

$$(\boldsymbol{I}_{NT} - \boldsymbol{P}_D)\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \\ \vdots \\ \boldsymbol{Z}_N \end{pmatrix} - \begin{pmatrix} \boldsymbol{\ell}_T\bar{\boldsymbol{z}}_1' \\ \boldsymbol{\ell}_T\bar{\boldsymbol{z}}_2' \\ \vdots \\ \boldsymbol{\ell}_T\bar{\boldsymbol{z}}_N' \end{pmatrix},$$

with the t^{th} observation in the i^{th} block being $(\boldsymbol{z}_{it} - \bar{\boldsymbol{z}}_i)'$, the deviation of $\boldsymbol{z}_{it}$ from its group average. This shows that the OLS estimator (4.22) can be obtained by regressing

$y_{it} - \bar{\boldsymbol{y}}_i$ on $\boldsymbol{z}_{it} - \bar{\boldsymbol{z}}_i$, $i = 1, \ldots, N$, and $t = 1, \ldots, T$. That is,

$$\begin{aligned}\hat{\boldsymbol{b}}_{NT} &= \left(\sum_{i=1}^{N}(\boldsymbol{Z}_i' - \bar{\boldsymbol{z}}_i\boldsymbol{\ell}_T')(\boldsymbol{Z}_i - \boldsymbol{\ell}_T\bar{\boldsymbol{z}}_i')\right)^{-1}\left(\sum_{i=1}^{N}(\boldsymbol{Z}_i' - \bar{\boldsymbol{z}}_i\boldsymbol{\ell}_T')(\boldsymbol{y}_i - \boldsymbol{\ell}_T\bar{\boldsymbol{y}}_i)\right) \\ &= \left(\sum_{i=1}^{N}\sum_{t=1}^{T}(\boldsymbol{z}_{it} - \bar{\boldsymbol{z}}_i)(\boldsymbol{z}_{it} - \bar{\boldsymbol{z}}_i)'\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}(\boldsymbol{z}_{it} - \bar{\boldsymbol{z}}_i)(y_{it} - \bar{\boldsymbol{y}}_i)\right).\end{aligned} \tag{4.23}$$

This estimator is known as the *within-groups* estimator because it is based on the observations that are deviations from their own group averages.

Let $\hat{\boldsymbol{a}}_{NT}$ denote the OLS estimator of the vector $\boldsymbol{a}$ of individual effects. By the facts that

$$\boldsymbol{D}'\hat{\boldsymbol{y}} = \boldsymbol{D}'\boldsymbol{D}\hat{\boldsymbol{a}}_{NT} + \boldsymbol{D}'\boldsymbol{Z}\hat{\boldsymbol{b}}_{NT},$$

and that the OLS residual vector is orthogonal to $\boldsymbol{D}$, $\hat{\boldsymbol{a}}_{NT}$ can be computed as

$$\hat{\boldsymbol{a}}_{NT} = (\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{b}}_{NT}), \tag{4.24}$$

with the i^{th} element being

$$\hat{\boldsymbol{a}}_{NT,i} = \frac{1}{T}(\boldsymbol{\ell}_T'\boldsymbol{y}_i - \boldsymbol{\ell}_T'\boldsymbol{Z}_i\hat{\boldsymbol{b}}_{NT}) = \bar{\boldsymbol{y}}_i - \bar{\boldsymbol{z}}_i'\hat{\boldsymbol{b}}_{NT}.$$

When the classical conditions [A1] and [A2](i) hold for every equation in (4.20), we have

$$\mathbb{E}(\boldsymbol{y}_i) = \boldsymbol{\ell}_T a_{i,o} + \boldsymbol{Z}_i\boldsymbol{b}_o, \qquad i = 1, 2, \ldots, N.$$

Then, the OLS estimators $\hat{\boldsymbol{a}}_{NT}$ and $\hat{\boldsymbol{b}}_{NT}$ are unbiased for $\boldsymbol{a}_o$ and $\boldsymbol{b}_o$, where the i^{th} element of $\boldsymbol{a}_o$ is $a_{i,o}$. Similar to Section 4.6, to ensure the BLUE property of these estimators, it is also required that $\text{var}(\boldsymbol{y})$ is a scalar covariance matrix. This amounts to requiring that $\text{var}(\boldsymbol{y}_i) = \sigma_o^2\boldsymbol{I}_T$ for all i and that $\text{cov}(\boldsymbol{y}_i, \boldsymbol{y}_j) = \boldsymbol{0}$ for every $i \neq j$. Under these conditions, the variance-covariance matrix of $\hat{\boldsymbol{b}}_{NT}$ is

$$\begin{aligned}\text{var}(\hat{\boldsymbol{b}}_{NT}) &= \sigma_o^2[\boldsymbol{Z}'(\boldsymbol{I}_{NT} - \boldsymbol{P}_D)\boldsymbol{Z}]^{-1} \\ &= \sigma_o^2\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\boldsymbol{z}_{it} - \bar{\boldsymbol{z}}_i)(\boldsymbol{z}_{it} - \bar{\boldsymbol{z}}_i)'\right]^{-1}.\end{aligned}$$

It is also easy to verify that the covariance matrix of the i^{th} element of $\hat{\boldsymbol{a}}_{NT}$ is

$$\text{var}(\hat{\boldsymbol{a}}_{NT,i}) = \frac{1}{T}\sigma_o^2 + \bar{\boldsymbol{z}}_i'[\text{var}(\hat{\boldsymbol{b}}_{NT})]\bar{\boldsymbol{z}}_i; \tag{4.25}$$

see Exercise 4.8. The OLS estimator for the regression variance σ_o^2 is

$$\hat{\sigma}_{NT}^2 = \frac{1}{NT - N - k + 1} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \hat{\boldsymbol{a}}_{NT,i} - \boldsymbol{z}_{it}' \hat{\boldsymbol{b}}_{NT})^2.$$

Substituting $\hat{\sigma}_{NT}^2$ into the formulae of $\text{var}(\hat{\boldsymbol{b}}_{NT})$ and $\text{var}(\hat{\boldsymbol{a}}_{NT,i})$ we immediately obtain their OLS estimators. On the other hand, if $\text{var}(\boldsymbol{y}_i) = \sigma_i^2 \boldsymbol{I}_T$ so that the variances of y_{it} are constant within each group but different across groups, we have the problem of heteroskedasticity. If $\text{cov}(\boldsymbol{y}_i, \boldsymbol{y}_j) = \sigma_{ij} \boldsymbol{I}_T$ for some $i \neq j$, we have spatial correlations among groups, even though observations are serially uncorrelated. In both cases, the OLS estimators are no longer the BLUEs, and FGLS estimation is needed.

Observe that when [A1] and [A2](i) hold for every equation in (4.20),

$$\mathbb{E}(\bar{\boldsymbol{y}}_i) = a_{i,o} + \bar{\boldsymbol{Z}}_i \boldsymbol{b}_o, \qquad i = 1, 2, \ldots, N.$$

One may then expect to estimate the parameters from a specification based on group-averages. In particular, the estimator

$$\check{\boldsymbol{b}}_b = \left(\sum_{i=1}^{N} (\bar{\boldsymbol{z}}_i - \bar{\boldsymbol{z}})(\bar{\boldsymbol{z}}_i - \bar{\boldsymbol{z}})' \right)^{-1} \left(\sum_{i=1}^{N} (\bar{\boldsymbol{z}}_i - \bar{\boldsymbol{z}})(\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}}) \right) \tag{4.26}$$

is the OLS estimator computed from the following specification:

$$\bar{\boldsymbol{y}}_i = a + \bar{\boldsymbol{z}}_i' \boldsymbol{b} + \boldsymbol{e}_i, \qquad i = 1, \ldots, N. \tag{4.27}$$

This is so because the sample means of $\bar{\boldsymbol{y}}_i$ and $\bar{\boldsymbol{z}}_i$ are just their respective overall averages: $\bar{\boldsymbol{y}}$ and $\bar{\boldsymbol{z}}$. The estimator (4.26) is known as the *between-groups* estimator because it is based on the deviations of group averages from their overall averages. As shown in Exercise 4.9, the between-groups estimator is biased for $\boldsymbol{b}_o$ when fixed effects are present. This should not be surprising because, while there are $N + k - 1$ parameters in the fixed effects model, the specification (4.27) contains only N observations and only permits estimation of k parameters.

Consider also a specification ignoring individual effects:

$$\boldsymbol{y}_i = \boldsymbol{\ell}_T a + \boldsymbol{Z}_i \boldsymbol{b} + \boldsymbol{e}_i, \qquad i = 1, \ldots, N. \tag{4.28}$$

The OLS estimator of $\boldsymbol{b}$ is

$$\check{\boldsymbol{b}}_p = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} (\boldsymbol{z}_{it} - \bar{\boldsymbol{z}})(\boldsymbol{z}_{it} - \bar{\boldsymbol{z}})' \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} (\boldsymbol{z}_{it} - \bar{\boldsymbol{z}})(y_{it} - \bar{\boldsymbol{y}}) \right). \tag{4.29}$$

When [A1] and [A2](i) hold for every equation in (4.20), one can see that (4.28) is in effect a specification that omits $n-1$ relevant dummy variables. It follows that $\check{\boldsymbol{b}}_p$ is a biased estimator for $\boldsymbol{b}_o$. Alternatively, it can be shown that the estimator (4.29) is a weighted sum of the between- and within-groups estimators and hence known as the "pooled" estimator; see Exercise 4.10. The pooled estimator $\check{\boldsymbol{b}}_p$ is therefore biased because $\check{\boldsymbol{b}}_b$ is. These examples show that neither the between-groups estimator nor the pooled estimator is a proper choice for the fixed effects model.

4.7.2 Random Effects Model

Given the specification (4.20) that allows for individual effects:

$$\boldsymbol{y}_i = \boldsymbol{\ell}_T a_i + \boldsymbol{Z}_i \boldsymbol{b} + \boldsymbol{e}_i, \qquad i = 1, 2, \ldots, N,$$

we now treat a_i as random variables rather than parameters. Writing $a_i = a + u_i$ with $a = \mathbb{E}(a_i)$, the specification above can be expressed as

$$\boldsymbol{y}_i = \boldsymbol{\ell}_T a + \boldsymbol{Z}_i \boldsymbol{b} + \boldsymbol{\ell}_T u_i + \boldsymbol{e}_i, \qquad i = 1, 2, \ldots, N. \tag{4.30}$$

where $\boldsymbol{\ell}_T u_i$ and $\boldsymbol{e}_i$ form the error term. This specification differs from the fixed effects model in that the intercept terms do not vary across i. The presence of u_i also makes (4.30) different from the specification that does not allow for individual effects. Here, group heterogeneity due to individual effects is characterized by the random variable u_i and absorbed into the error term. Thus, (4.30) is known as the *random effects model*.

As far as regression coefficients are concerned, (4.30) and (4.28) are virtually the same. As such, the OLS estimator of $\boldsymbol{b}$ is just the pooled estimator $\check{\boldsymbol{b}}_p$. The OLS estimator of a is

$$\check{a}_p = \bar{\boldsymbol{y}} - \bar{\boldsymbol{z}}' \check{\boldsymbol{b}}_p.$$

If the classical conditions [A1] and [A2](i) hold for each equation such that

$$\mathbb{E}(\boldsymbol{y}_i) = \boldsymbol{\ell}_T a_o + \boldsymbol{Z}_i \boldsymbol{b}_o, \qquad i = 1, \ldots, N,$$

$\check{\boldsymbol{b}}_p$ and $\check{a}_p$ are unbiased for $\boldsymbol{b}_o$ and a_o. Note, however, that the pooled estimator would be biased if the individual effects were fixed, as shown in the preceding section.

When [A1] and [A2](i) hold, we can write

$$\boldsymbol{y}_i = \boldsymbol{\ell}_T a_o + \boldsymbol{Z}_i \boldsymbol{b}_o + \check{\boldsymbol{\epsilon}}_i, \tag{4.31}$$

where $\check{\boldsymbol{\epsilon}}_i = \boldsymbol{\ell}_T u_i + \boldsymbol{\epsilon}_i$. That is, $\check{\boldsymbol{\epsilon}}_i$ contains two components: the random effects $\boldsymbol{\ell}_T u_i$ and the disturbance $\boldsymbol{\epsilon}_i$ which exists even when there is no random effect. Thus,

$$\operatorname{var}(\boldsymbol{y}_i) = \sigma_u^2 \boldsymbol{\ell}_T \boldsymbol{\ell}_T' + \operatorname{var}(\boldsymbol{\epsilon}_i) + 2\operatorname{cov}(\boldsymbol{\ell}_T u_i, \boldsymbol{\epsilon}_i),$$

where σ_u^2 is $\operatorname{var}(u_i)$. As the first term on the right-hand side above is a full matrix, $\operatorname{var}(\boldsymbol{y}_i)$ is not a scalar covariance matrix in general. It follows that $\check{\boldsymbol{b}}_p$ and $\check{a}_p$ are not the BLUEs.

To perform FGLS estimation, more conditions on $\operatorname{var}(\boldsymbol{y}_i)$ are needed. If $\operatorname{var}(\boldsymbol{\epsilon}_i) = \sigma_o^2 \boldsymbol{I}_T$ and $\mathbb{E}(u_i \boldsymbol{\epsilon}_i) = \boldsymbol{0}$, we obtain a simpler form of $\operatorname{var}(\boldsymbol{y}_i)$:

$$\boldsymbol{S}_o := \operatorname{var}(\boldsymbol{y}_i) = \sigma_u^2 \boldsymbol{\ell}_T \boldsymbol{\ell}_T' + \sigma_o^2 \boldsymbol{I}_T.$$

Under additional conditions that $\mathbb{E}(u_i u_j) = 0$, $E(u_i \boldsymbol{\epsilon}_j) = \boldsymbol{0}$ and $\mathbb{E}(\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_j) = \boldsymbol{0}$ for all $i \neq j$, we have $\operatorname{cov}(\boldsymbol{y}_i, \boldsymbol{y}_j) = \boldsymbol{0}$. Hence, $\operatorname{var}(\boldsymbol{y})$ simplifies to a block diagonal matrix:

$$\boldsymbol{\Sigma}_o := \operatorname{var}(\boldsymbol{y}) = \boldsymbol{I}_N \otimes \boldsymbol{S}_o.$$

It can be verified that the desired transformation matrix for GLS estimation is $\boldsymbol{\Sigma}_o^{-1/2} = I_N \otimes \boldsymbol{S}_o^{-1/2}$, where

$$\boldsymbol{S}_o^{-1/2} = \boldsymbol{I}_T - \frac{c}{T} \boldsymbol{\ell}_T \boldsymbol{\ell}_T',$$

and $c = 1 - \sigma_o^2/(T\sigma_u^2 + \sigma_o^2)^{1/2}$. Transformed data are $\boldsymbol{S}_o^{-1/2} \boldsymbol{y}_i$ and $\boldsymbol{S}_o^{-1/2} \boldsymbol{Z}_i$, $i = 1, \ldots, N$, and their t^{th} elements are, respectively, $y_{it} - c\bar{\boldsymbol{y}}_i$ and $\boldsymbol{z}_{it} - c\bar{\boldsymbol{z}}_i$. If $\sigma_o^2 = 0$ so that the disturbances $\boldsymbol{\epsilon}_i$ are absent, we have $c = 1$, so that

$$\boldsymbol{\Sigma}_o^{-1/2} = \boldsymbol{I}_N \otimes (\boldsymbol{I}_T - \boldsymbol{\ell}_T \boldsymbol{\ell}_T'/T) = \boldsymbol{I}_{NT} - \boldsymbol{P}_D,$$

as in the fixed effects model. Consequently, the GLS estimator of $\boldsymbol{b}$ is nothing but the within-groups estimator (4.22). It can be shown that the GLS estimator is also a weighted average of the within- and between-groups estimators.

To compute the FGLS estimator, we must estimate σ_u^2 and σ_o^2. Pre-multiplying (4.31) by $\boldsymbol{\ell}_T'/T$ yields

$$\bar{\boldsymbol{y}}_i = a_o + \bar{\boldsymbol{z}}_i' \boldsymbol{b}_o + u_i + \bar{\boldsymbol{\epsilon}}_i.$$

Taking the difference of $\boldsymbol{y}_i$ and $\bar{\boldsymbol{y}}_i$ we have

$$\boldsymbol{y}_i - \boldsymbol{\ell}_T \bar{\boldsymbol{y}}_i = (\boldsymbol{Z}_i - \boldsymbol{\ell}_T \bar{\boldsymbol{z}}_i') \boldsymbol{b}_o + (\boldsymbol{\epsilon}_i - \boldsymbol{\ell}_T \bar{\boldsymbol{\epsilon}}_i),$$

which does not involve u_i. This suggests that, even when random effects are present, we can also estimate $\boldsymbol{b}_o$ from the specification

$$\boldsymbol{y}_i - \boldsymbol{\ell}_T\bar{\boldsymbol{y}}_i = (\boldsymbol{Z}_i - \boldsymbol{\ell}_T\bar{\boldsymbol{z}}_i')\boldsymbol{b} + \boldsymbol{e}_i, \qquad i = 1, \ldots, N.$$

It is readily seen that the OLS estimator based on this specification is the within-groups estimator $\hat{\boldsymbol{b}}_{NT}$. As u_i have been eliminated, we can estimate σ_o^2, the variance of ϵ_{it}, by

$$\hat{\sigma}_\epsilon^2 = \frac{1}{NT - N - k + 1}\sum_{i=1}^{N}\sum_{i=1}^{T}[(y_{it} - \bar{\boldsymbol{y}}_i) - (\boldsymbol{z}_{it} - \bar{\boldsymbol{z}}_i)'\hat{\boldsymbol{b}}_{NT}]^2,$$

which is also the variance estimator in the fixed effects model.

By (4.31),

$$\bar{\boldsymbol{y}}_i = a_o + \bar{\boldsymbol{z}}_i'\boldsymbol{b}_o + u_i + \bar{\boldsymbol{\epsilon}}_i.$$

This suggests that $\boldsymbol{b}_o$ can be estimated from the specification based on group averages:

$$\bar{\boldsymbol{y}}_i = a + \bar{\boldsymbol{z}}_i'\boldsymbol{b} + \boldsymbol{e}_i, \qquad i = 1, \ldots, N.$$

This specification is the same as (4.27) so that the OLS estimator of $\boldsymbol{b}$ is the between-groups estimator $\hat{\boldsymbol{b}}_b$. The resulting OLS residuals are

$$\check{e}_i = (\bar{\boldsymbol{y}}_i - \bar{\boldsymbol{y}}) - (\bar{\boldsymbol{z}}_i - \bar{\boldsymbol{z}})'\check{\boldsymbol{b}}_b, \qquad i = 1, \ldots, N.$$

We can estimate

$$\mathrm{var}(u_i + \bar{\boldsymbol{\epsilon}}_i) = \sigma_u^2 + \sigma_o^2/T$$

by $\sum_{i=1}^{N}\check{e}_i^2/(N-k)$, from which the estimator for σ_u^2 can be calculated as

$$\hat{\sigma}_u^2 = \frac{1}{N-k}\sum_{i=1}^{N}\check{e}_i^2 - \frac{\hat{\sigma}_\epsilon^2}{T}.$$

With these two estimators $\hat{\sigma}_u^2$ and $\hat{\sigma}_\epsilon^2$ we can construct the transformation matrix for the FGLS estimator, $\hat{S}^{-1/2}$. It is clear that the FGLS estimator is, again, a very complex function of $\boldsymbol{y}$.

4.8 Limitations of the FGLS Method

In this chapter we relax only the classical condition [A2](ii) while maintaining [A1] and [A2](i). The limitations of [A1] and [A2](i) discussed in Chapter 3.7 therefore still

exist. In particular, stochastic regressors and nonlinear specifications are excluded in the present context.

Although the GLS and FGLS methods are designed to improve on estimation efficiency when there is a non-scalar covariance matrix $\boldsymbol{\Sigma}_o$, they also create further difficulties. First, the GLS estimator is usually not available, except in some exceptional cases. Second, a convenient FGLS estimator is available at the expense of more conditions on $\boldsymbol{\Sigma}_o$. If these simplifying conditions are incorrectly imposed, the resulting FGLS estimator may perform poorly. Third, the finite-sample properties of the FGLS estimator are typically unknown. In general, we do not know if an FGLS estimator is unbiased, nor do we know its efficiency relative to the OLS estimator and its exact distribution. It is therefore difficult to draw statistical inferences from FGLS estimation results.

Exercises

4.1 Given the linear specification $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, suppose that the conditions [A1] and [A2](ii) hold and that $\mathrm{var}(\boldsymbol{y}) = \boldsymbol{\Sigma}_o$. If the matrix $\boldsymbol{X}$ contains k eigenvectors of $\boldsymbol{\Sigma}_o$ which are normalized to unit length. What are the resulting $\hat{\boldsymbol{\beta}}_T$ and $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$? Explain your result.

4.2 Show that R^2_{GLS} need not be bounded between zero and one.

4.3 Given the linear specification $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, suppose that the conditions [A1] and [A2](i) hold and that $\mathrm{var}(\boldsymbol{y}) = \boldsymbol{\Sigma}_o$. Show directly that

$$\mathrm{var}(\hat{\boldsymbol{\beta}}_T) - \mathrm{var}(\hat{\boldsymbol{\beta}}_{\mathrm{GLS}})$$

is a positive semi-definite matrix.

4.4 Suppose that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_o + \boldsymbol{\epsilon}$ and the elements of $\boldsymbol{\epsilon}$ are $\epsilon_t = \alpha_1\epsilon_{t-1} + u_t$, where $\alpha_1 = 1$ and $\{u_t\}$ is a white noise with mean zero and variance σ_u^2. What are the properties of ϵ_t? Is $\{\epsilon_t\}$ still weakly stationary?

4.5 Suppose that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_o + \boldsymbol{\epsilon}$ and the elements of $\boldsymbol{\epsilon}$ are $\epsilon_t = u_t + \alpha_1 u_{t-1}$, where $|\alpha_1| < 1$ and $\{u_t\}$ is a white noise with mean zero and variance σ_u^2. Calculate the variance, autocovariances, and autocorrelations of ϵ_t and compare them with those of AR(1) disturbances.

4.6 Let y_t denote investment expenditure that is determined by expected earning x_t^*:

$$y_t = a_o + b_o x_t^* + u_t.$$

When x_t^* is adjusted adaptively:

$$x_t^* = x_{t-1}^* + (1 - \lambda_o)(x_t - x_{t-1}^*), \qquad 0 < \lambda_o < 1,$$

show that y_t can be represented by a model with a lagged dependent variable and moving average disturbances.

4.7 Given the SUR specification (4.17), show that the GLS estimator is the same as the OLS estimator when $\boldsymbol{X}_i = \boldsymbol{X}_0$ for all i. Give an intuitive explanation of this result.

4.8 Given the specification (4.20), suppose that [A1] and [A2](i) hold for each group equation, $\text{var}(\boldsymbol{y}_i) = \sigma_o^2 \boldsymbol{I}_T$ and $\text{cov}(\boldsymbol{y}_i, \boldsymbol{y}_j) = \boldsymbol{0}$ for $i \neq j$. Prove (4.25).

4.9 Given the specification (4.27), suppose that [A1] and [A2](i) hold for each group equation. Find the expected value of the between-groups estimator (4.26).

4.10 Given the specification (4.28), show that the pooled estimator (4.29) is a weighted sum of the within- and between-groups estimators. What is its expected value when [A1] and [A2](i) hold for each group equation.

References

Amemiya, Takeshi (1985). *Advanced Econometrics*, Cambridge, MA: Harvard University Press.

Farebrother, R. W. (1980). The Durbin-Watson test for serial correlation when there is no intercept in the regression, *Econometrica*, **48**, 1553–1563.

Goldberger, Arthur S. (1991). *A Course in Econometrics*, Cambridge, MA: Harvard University Press.

Greene, William H. (2000). *Econometric Analysis*, Fourth ed., New York, NY: Macmillan.

Wallis, K. F. (1972). Testing for fourth order autocorrelation in quarterly regression equations, *Econometrica*, **40**, 617–636.

Chapter 5

Probability Theory

The purpose of this chapter is to summarize some important concepts and results in probability theory to be used subsequently. We formally define random variables and moments (unconditional and conditional) under a measure-theoretic framework. Our emphasis is on important limiting theorems, such as the law of large numbers and central limit theorem, which play a crucial role in the asymptotic analysis of many econometric estimators and tests. Davidson (1994) provides a complete and thorough treatment of the topics in this chapter; see also Bierens (1994), Gallant (1997) and White (1984) for a concise coverage. Many results here are taken freely from these references. The readers may also consult other real analysis and probability textbooks for related topics.

5.1 Probability Space and Random Variables

5.1.1 Probability Space

The probability space associated with a random experiment is determined by three components: the outcome space Ω, a collection of events (subsets of Ω) $\mathcal{F}$, and a probability measure assigned to the elements in $\mathcal{F}$. Given the subset A of Ω, its complement is $A^c = \{\omega \in \Omega : \omega \notin A\}$.

In the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{F}$ is a σ-algebra (σ-field) in the sense that it satisfies the following requirements:

1. $\Omega \in \mathcal{F}$;
2. if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$;
3. if $A_1, A_2, \ldots$ are in $\mathcal{F}$, then $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$.

The first and second properties imply that $\Omega^c = \emptyset$ is also in $\mathcal{F}$. Combining the second and third properties we have from de Morgan's law that

$$\left(\bigcup_{n=1}^{\infty} A_n\right)^c = \bigcap_{n=1}^{\infty} A_n^c \in \mathcal{F}.$$

A σ-algebra is thus closed under complementation, countable union and countable intersection.

The probability measure $\mathbb{P} : \mathcal{F} \mapsto [0,1]$ is a real-valued set function satisfying the following axioms:

1. $\mathbb{P}(\Omega) = 1$;
2. $\mathbb{P}(A) \geq 0$ for all $A \in \mathcal{F}$;
3. if $A_1, A_2, \ldots \in \mathcal{F}$ are disjoint, then $\mathbb{P}(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$.

From these axioms we easily deduce that $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, $\mathbb{P}(A) \leq \mathbb{P}(B)$ if $A \subseteq B$, and

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Moreover, if $\{A_n\}$ is an increasing (decreasing) sequence in $\mathcal{F}$ with the limiting set A, then $\lim_n \mathbb{P}(A_n) = \mathbb{P}(A)$.

Let $\mathcal{C}$ be a collection of subsets of Ω. The intersection of all the σ-algebras that contain $\mathcal{C}$ is the smallest σ-algebra containing $\mathcal{C}$; see Exercise 5.1. This σ-algebra is referred to as the σ-algebra generated by $\mathcal{C}$, denoted as $\sigma(\mathcal{C})$. When $\Omega = \mathbb{R}$, the *Borel field* is the σ-algebra generated by all open intervals (a, b) in $\mathbb{R}$. Note that open intervals, closed intervals $[a, b]$, half-open intervals $(a, b]$ or half lines $(-\infty, b]$ can be obtained from each other by taking complement, union and/or intersection. For example,

$$(a, b] = \bigcap_{n=1}^{\infty} \left(a,\ b + \frac{1}{n}\right), \qquad (a, b) = \bigcup_{n=1}^{\infty} \left(a, b - \frac{1}{n}\right].$$

Thus, the collection of all closed intervals (half-open intervals, half lines) generates the same Borel field. As such, open intervals, closed intervals, half-open intervals and half lines are also known as *Borel sets*. The Borel field on $\mathbb{R}^d$, denoted as $\mathcal{B}^d$, is generated by all open hypercubes:

$$(a_1, b_1) \times (a_2, b_2) \times \cdots \times (a_d, b_d).$$

Equivalently, $\mathcal{B}^d$ can be generated by all closed (half-open) hypercubes, or by

$$(-\infty, b_1] \times (-\infty, b_2] \times \cdots \times (-\infty, b_d].$$

5.1.2 Random Variables

Let $\mathcal{B}$ denote the Borel field on $\mathbb{R}$. A random variable z is a function $z\colon \Omega \mapsto \mathbb{R}$ such that for every $B \in \mathcal{B}$, the *inverse image* of B under z is in $\mathcal{F}$, i.e.,

$$z^{-1}(B) = \{\omega\colon z(\omega) \in B\} \in \mathcal{F}.$$

We also say that z is a $\mathcal{F}/\mathcal{B}$-measurable (or simply $\mathcal{F}$-measurable) function. A $\mathbb{R}^d$-valued random variable $\boldsymbol{z}$ is a function $\boldsymbol{z}\colon \Omega \to \mathbb{R}^d$ that is $\mathcal{F}/\mathcal{B}^d$-measurable. Given the random vector $\boldsymbol{z}$, its inverse images $\boldsymbol{z}^{-1}(B)$ form a σ-algebra, denoted as $\sigma(\boldsymbol{z})$. It can be shown that $\sigma(\boldsymbol{z})$ is the smallest σ-algebra contained in $\mathcal{F}$ such that $\boldsymbol{z}$ is measurable. We usually interpret $\sigma(\boldsymbol{z})$ as the set containing all the information associated with $\boldsymbol{z}$.

A function $g\colon \mathbb{R} \mapsto \mathbb{R}$ is said to be $\mathcal{B}$-measurable or *Borel measurable* if

$$\{\zeta \in \mathbb{R}\colon g(\zeta) \le b\} \in \mathcal{B}.$$

If z is a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$, then $g(z)$ is also a random variable defined on the same probability space provided that g is Borel measurable. Note that the functions we usually encounter are indeed Borel measurable; non-measurable functions are very exceptional and hence are not of general interest. Similarly, for the d-dimensional random vector $\boldsymbol{z}$, $g(\boldsymbol{z})$ is a random variable provided that g is $\mathcal{B}^d$-measurable.

Recall from Section 2.1 that the joint distribution function of $\boldsymbol{z}$ is the non-decreasing, right-continuous function $F_{\boldsymbol{z}}$ such that for $\boldsymbol{\zeta} = (\zeta_1 \ \cdots \ \zeta_d)' \in \mathbb{R}^d$,

$$F_{\boldsymbol{z}}(\boldsymbol{\zeta}) = \mathbb{P}\{\omega \in \Omega\colon z_1(\omega) \le \zeta_1, \ldots, z_d(\omega) \le \zeta_d\},$$

with

$$\lim_{\zeta_1 \to -\infty, \ldots, \zeta_d \to -\infty} F_{\boldsymbol{z}}(\boldsymbol{\zeta}) = 0, \qquad \lim_{\zeta_1 \to \infty, \ldots, \zeta_d \to \infty} F_{\boldsymbol{z}}(\boldsymbol{\zeta}) = 1.$$

The marginal distribution function of the i^{th} component of $\boldsymbol{z}$ is such that

$$F_{z_i}(\zeta_i) = \mathbb{P}\{\omega \in \Omega\colon z_i(\omega) \le \zeta_i\} = F_{\boldsymbol{z}}(\infty, \ldots, \infty, \zeta_i, \infty, \ldots, \infty).$$

Note that while $\mathbb{P}$ is a set function defined on $\mathcal{F}$, the distribution function of $\boldsymbol{z}$ is a point function defined on $\mathbb{R}^d$.

Two random variables y and z are said to be (pairwise) *independent* if, and only if, for any Borel sets B_1 and B_2,

$$\mathbb{P}(y \in B_1 \text{ and } z \in B_2) = \mathbb{P}(y \in B_1)\ \mathbb{P}(z \in B_2).$$

This immediately leads to the standard definition of independence: y and z are independent if, and only if, their joint distribution is the product of their marginal distributions, as in Section 2.1. A sequence of random variables $\{z_i\}$ is said to be *totally independent* if

$$\mathbb{P}\left(\bigcap_{\text{all } i}\{z_i \in B_i\}\right) = \prod_{\text{all } i}\mathbb{P}(z_i \in B_i),$$

for any Borel sets B_i. In what follows, a totally independent sequence will be referred to an independent sequence or a sequence of independent variables for convenience. For an independent sequence, we have the following generalization of Lemma 2.1.

Lemma 5.1 *Let $\{z_i\}$ be a sequence of independent random variables and h_i, $i = 1, 2, \ldots$, be Borel-measurable functions. Then $\{h_i(z_i)\}$ is also a sequence of independent random variables.*

5.1.3 Moments and Norms

The expectation of the i^{th} element of $\boldsymbol{z}$ is

$$\mathbb{E}(z_i) = \int_\Omega z_i(\omega)\,\mathrm{d}\,\mathbb{P}(\omega),$$

where the right-hand side is a Lebesgue integral. In view of the distribution function defined above, a change of ω causes the realization of $\boldsymbol{z}$ to change so that

$$\mathbb{E}(z_i) = \int_{\mathbb{R}^d} \zeta_i\,\mathrm{d}F_{\boldsymbol{z}}(\boldsymbol{\zeta}) = \int_{\mathbb{R}} \zeta_i\,\mathrm{d}F_{z_i}(\zeta_i),$$

where F_{z_i} is the marginal distribution function of the i^{th} component of $\boldsymbol{z}$, as defined in Section 2.2. For the Borel measurable function g of $\boldsymbol{z}$,

$$\mathbb{E}[g(\boldsymbol{z})] = \int_\Omega g(\boldsymbol{z}(\omega))\,\mathrm{d}\,\mathbb{P}(\omega) = \int_{\mathbb{R}^d} g(\boldsymbol{\zeta})\,\mathrm{d}F_{\boldsymbol{z}}(\boldsymbol{\zeta}).$$

Other moments, such as variance and covariance, can also be defined as Lebesgue integrals with respect to the probability measure; see Section 2.2.

A function g is said to be *convex* on a set S if for any $a \in [0,1]$ and any x, y in S,

$$g\big(ax + (1-a)y\big) \le ag(x) + (1-a)g(y);$$

g is *concave* on S if the inequality above is reversed. For example, $g(x) = x^2$ is convex, and $g(x) = \log x$ for $x > 0$ is concave. The result below is concerned with convex (concave) transformations.

Lemma 5.2 (Jensen) *For the Borel measurable function g that is convex on the support of the integrable random variable z, suppose that $g(z)$ is also integrable. Then,*

$$g(\mathbb{E}(z)) \leq \mathbb{E}[g(z)];$$

the inequality reverses if g is concave.

For the random variable z with finite p^{th} moment, let $\|z\|_p = [\mathbb{E}(z^p)]^{1/p}$ denote its L_p-norm. Also define the *inner product* of two square integrable random variables z_i and z_j as their cross moment:

$$\langle z_i, z_j \rangle = \mathbb{E}(z_i z_j).$$

Then, L_2-norm can be obtained from the inner product as $\|z_i\|_2 = \langle z_i, z_i \rangle^{1/2}$. It is easily seen that for any $c > 0$ and $p > 0$,

$$c^p \, \mathbb{P}(|z| \geq c) = c^p \int \mathbf{1}_{\{\zeta : |\zeta| \geq c\}} \, \mathrm{d}F_z(\zeta) \leq \int_{\{\zeta : |\zeta| \geq c\}} |\zeta|^p \, \mathrm{d}F_z(\zeta) \leq \mathbb{E}\,|z|^p,$$

where $\mathbf{1}_{\{\zeta : |\zeta| \geq c\}}$ is the indicator function which equals one if $|\zeta| \geq c$ and equals zero otherwise. This establishes the following result.

Lemma 5.3 (Markov) *Let z be a random variable with finite $p^{\,th}$ moment. Then,*

$$\mathbb{P}(|z| \geq c) \leq \frac{\mathbb{E}\,|z|^p}{c^p},$$

where c is a positive real number.

For $p = 2$, Lemma 5.3 is also known as the *Chebyshev inequality*. If c is small such that $\mathbb{E}\,|z|^p / c^p > 1$, Markov's inequality is trivial. When c tends to infinity, the probability that z assumes very extreme values will be vanishing at the rate c^{-p}.

Another useful result in probability theory is stated below without proof.

Lemma 5.4 (Hölder) *Let y be a random variable with finite $p^{\,th}$ moment ($p > 1$) and z a random variable with finite $q^{\,th}$ moment ($q = p/(p-1)$). Then,*

$$\mathbb{E}\,|yz| \leq \|y\|_p \, \|z\|_q.$$

For $p = 2$, we have $\mathbb{E}\,|yz| \leq \|y\|_2 \, \|z\|_2$. By noting that $|\,\mathbb{E}(yz)| < \mathbb{E}\,|yz|$, we immediately have the next result; cf. Lemma 2.3.

Lemma 5.5 (Cauchy-Schwartz) *Let y and z be two square integrable random variables. Then,*

$$|\,\mathrm{I\!E}(yz)| \leq \|y\|_2 \, \|z\|_2.$$

Let $y = 1$ and $x = z^p$. Then for $q > p$ and $r = q/p$, Hölder's inequality also ensures that

$$\mathrm{I\!E}\,|z^p| \leq \|x\|_r \, \|y\|_{r/(r-1)} = [\mathrm{I\!E}(z^{pr})]^{1/r} = [\mathrm{I\!E}(z^q)]^{p/q}.$$

This shows that when a random variable has finite $q^{\,\text{th}}$ moment, it must also have finite $p^{\,\text{th}}$ moment for any $p < q$, as stated below.

Lemma 5.6 (Liapunov) *Let z be a random variable with finite $q^{\,th}$ moment. Then for $p < q$, $\|z\|_p \leq \|z\|_q$.*

The inequality below states that the L_p-norm of a finite sum is less than the sum of individual L_p-norms.

Lemma 5.7 (Minkowski) *Let z_i, $i = 1, \ldots, n$, be random variables with finite $p^{\,th}$ moment ($p \geq 1$). Then,*

$$\left\| \sum_{i=1}^{n} z_i \right\|_p \leq \sum_{i=1}^{n} \|z_i\|_p.$$

When there are only two random variables in the sum, this is just the *triangle inequality* for L_p-norms; see also Exercise 5.4.

5.2 Conditional Distribution and Moments

Given two events A and B in $\mathcal{F}$, if it is known that B has occurred, the outcome space is restricted to B, so that the outcomes of A must be in $A \cap B$. The likelihood of A is thus characterized by the conditional probability

$$\mathrm{I\!P}(A \mid B) = \mathrm{I\!P}(A \cap B)/\,\mathrm{I\!P}(B),$$

for $\mathrm{I\!P}(B) \neq 0$. It can be shown that $\mathrm{I\!P}(\cdot|B)$ satisfies the axioms for probability measures; see Exerise 5.5. This concept is readily extended to construct *conditional density function* and *conditional distribution function*.

5.2.1 Conditional Distributions

Let $\boldsymbol{y}$ and $\boldsymbol{z}$ denote two integrable random vectors such that $\boldsymbol{z}$ has the density function $f_{\boldsymbol{z}}$. For $f_{\boldsymbol{z}}(\boldsymbol{\zeta}) \neq 0$, define the conditional density function of $\boldsymbol{z}$ given $\boldsymbol{y} = \boldsymbol{\eta}$ as

$$f_{\boldsymbol{z}|\boldsymbol{y}}(\boldsymbol{\zeta} \mid \boldsymbol{y} = \boldsymbol{\eta}) = \frac{f_{\boldsymbol{z},\boldsymbol{y}}(\boldsymbol{\zeta}, \boldsymbol{\eta})}{f_{\boldsymbol{y}}(\boldsymbol{\eta})},$$

which is clearly non-negative whenever it is defined. This function also integrates to one on $\mathbb{R}^d$ because

$$\int_{\mathbb{R}^d} f_{\boldsymbol{z}|\boldsymbol{y}}(\boldsymbol{\zeta} \mid \boldsymbol{y} = \boldsymbol{\eta})\, \mathrm{d}\boldsymbol{\zeta} = \frac{1}{f_{\boldsymbol{y}}(\boldsymbol{\eta})} \int_{\mathbb{R}^d} f_{\boldsymbol{z},\boldsymbol{y}}(\boldsymbol{\zeta}, \boldsymbol{\eta})\, \mathrm{d}\boldsymbol{\zeta} = \frac{1}{f_{\boldsymbol{y}}(\boldsymbol{\eta})} f_{\boldsymbol{y}}(\boldsymbol{\eta}) = 1.$$

Thus, $f_{\boldsymbol{z}|\boldsymbol{y}}$ is a legitimate density function. For example, the bivariate density function of two random variables z and y forms a surface on the zy-plane. By fixing $y = \eta$, we obtain a cross section (slice) under this surface. Dividing the joint density by the marginal density $f_y(\eta)$ amounts to adjusting the height of this slice so that the resulting area integrates to one.

Given the conditional density function $f_{\boldsymbol{z}|\boldsymbol{y}}$, we have for $A \in \mathcal{B}^d$,

$$\mathbb{P}(\boldsymbol{z} \in A \mid \boldsymbol{y} = \boldsymbol{\eta}) = \int_A f_{\boldsymbol{z}|\boldsymbol{y}}(\boldsymbol{\zeta} \mid \boldsymbol{y} = \boldsymbol{\eta})\, \mathrm{d}\boldsymbol{\zeta}.$$

Note that this conditional probability is defined even when $\mathbb{P}(\boldsymbol{y} = \boldsymbol{\eta})$ may be zero. In particular, when

$$A = (-\infty, \zeta_1] \times \cdots \times (-\infty, \zeta_d],$$

we obtain the *conditional distribution* function:

$$F_{\boldsymbol{z}|\boldsymbol{y}}(\boldsymbol{\zeta} \mid \boldsymbol{y} = \boldsymbol{\eta}) = \mathbb{P}(z_1 \leq \zeta_1, \ldots, z_d \leq \zeta_d \mid \boldsymbol{y} = \boldsymbol{\eta}).$$

When $\boldsymbol{z}$ and $\boldsymbol{y}$ are independent, the conditional density (distribution) simply reduces to the unconditional density (distribution).

5.2.2 Conditional Moments

Analogous to unconditional expectation, the *conditional expectation* of the integrable random variable z_i given the information $\boldsymbol{y} = \boldsymbol{\eta}$ is

$$\mathbb{E}(z_i \mid \boldsymbol{y} = \boldsymbol{\eta}) = \int_{\mathbb{R}} \zeta_i\, \mathrm{d}F_{\boldsymbol{z}|\boldsymbol{y}}(\zeta_i \mid \boldsymbol{y} = \boldsymbol{\eta});$$

$\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y} = \boldsymbol{\eta})$ is defined elementwise. By allowing $\boldsymbol{y}$ to vary across all possible values $\boldsymbol{\eta}$, we obtain the conditional expectation function $\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y})$ whose realization depends on $\boldsymbol{\eta}$, the realization of $\boldsymbol{y}$. Thus, $\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y})$ is a function of $\boldsymbol{y}$ and hence a random vector.

More generally, we can take a suitable σ-algebra as a conditioning set and define $\mathbb{E}(\boldsymbol{z} \mid \mathcal{G})$, where $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}$. Similar to the discussion above, $\mathbb{E}(\boldsymbol{z} \mid \mathcal{G})$ varies with the occurrence of each $G \in \mathcal{G}$. Specifically, for the integrable random vector $\boldsymbol{z}$, $\mathbb{E}(\boldsymbol{z} \mid \mathcal{G})$ is the $\mathcal{G}$-measurable random variable satisfying

$$\int_G \mathbb{E}(\boldsymbol{z} \mid \mathcal{G}) \, \mathrm{d}\,\mathbb{P} = \int_G \boldsymbol{z} \, \mathrm{d}\,\mathbb{P},$$

for all $G \in \mathcal{G}$. By setting $\mathcal{G} = \sigma(\boldsymbol{y})$, the σ-algebra generated by $\boldsymbol{y}$, we can write

$$\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y}) = \mathbb{E}[\boldsymbol{z} \mid \sigma(\boldsymbol{y})],$$

which is interpreted as the expectation of $\boldsymbol{z}$ given all the information associated with $\boldsymbol{y}$. Note that the unconditional expectation $\mathbb{E}(\boldsymbol{z})$ can be viewed as the expectation of $\boldsymbol{z}$ conditional on the trivial σ-algebra $\{\Omega, \emptyset\}$, i.e., the smallest σ-algebra that contains no extra information from any random vectors.

Similar to unconditional expectations, conditional expectations are monotonic: if $z \geq x$ with probability one, then $\mathbb{E}(z \mid \mathcal{G}) \geq \mathbb{E}(x \mid \mathcal{G})$ with probability one; in particular, if $z \geq 0$ with probability one, then $\mathbb{E}(z \mid \mathcal{G}) \geq 0$ with probability one. Moreover, If $\boldsymbol{z}$ is independent of $\boldsymbol{y}$, then $\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y}) = \mathbb{E}(\boldsymbol{z})$. For example, if $\boldsymbol{z}$ is a constant vector $\boldsymbol{c}$ which is independent of any random variable, then $\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y}) = \boldsymbol{c}$. The linearity result below is analogous to Lemma 2.2 for unconditional expectations.

Lemma 5.8 *Let $\boldsymbol{z}$ $(d \times 1)$ and $\boldsymbol{y}$ $(c \times 1)$ be integrable random vectors and $\boldsymbol{A}$ $(n \times d)$ and $\boldsymbol{B}$ $(n \times c)$ be non-stochastic matrices. Then with probability one,*

$$\mathbb{E}(\boldsymbol{A}\boldsymbol{z} + \boldsymbol{B}\boldsymbol{y} \mid \mathcal{G}) = \boldsymbol{A}\,\mathbb{E}(\boldsymbol{z} \mid \mathcal{G}) + \boldsymbol{B}\,\mathbb{E}(\boldsymbol{y} \mid \mathcal{G}).$$

If $\boldsymbol{b}$ $(n \times 1)$ is a non-stochastic vector, $\mathbb{E}(\boldsymbol{A}\boldsymbol{z} + \boldsymbol{b} \mid \mathcal{G}) = \boldsymbol{A}\,\mathbb{E}(\boldsymbol{z} \mid \mathcal{G}) + \boldsymbol{b}$ with probability one.

From the definition of conditional expectation, we immediately have

$$\int_\Omega \mathbb{E}(\boldsymbol{z} \mid \mathcal{G}) \, \mathrm{d}\,\mathbb{P} = \int_\Omega \boldsymbol{z} \, \mathrm{d}\,\mathbb{P};$$

that is, $\mathbb{E}[\mathbb{E}(\boldsymbol{z} \mid \mathcal{G})] = \mathbb{E}(\boldsymbol{z})$. This is known as the *law of iterated expectations*. As $\mathbb{E}(\boldsymbol{z})$ is also the conditional expectation with respect to the trivial (smallest) σ-algebra,

the equality above suggests that if conditional expectations are taken sequentially with respect to different σ-algebras, only the one with respect to a smaller σ-algebra matters. For example, for k random vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$,

$$\mathbb{E}[\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_k) \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{k-1}] = \mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{k-1}).$$

A more general result is given below; see Exercise 5.6.

Lemma 5.9 (Law of Iterated Expectations) *Let $\mathcal{G}$ and $\mathcal{H}$ be two sub-σ-algebras of $\mathcal{F}$ such that $\mathcal{G} \subseteq \mathcal{H}$. Then for the integrable random vector $\boldsymbol{z}$,*

$$\mathbb{E}[\mathbb{E}(\boldsymbol{z} \mid \mathcal{H}) \mid \mathcal{G}] = \mathbb{E}[\mathbb{E}(\boldsymbol{z} \mid \mathcal{G}) \mid \mathcal{H}] = \mathbb{E}(\boldsymbol{z} \mid \mathcal{G});$$

in particular, $\mathbb{E}[\mathbb{E}(\boldsymbol{z} \mid \mathcal{G})] = \mathbb{E}(\boldsymbol{z})$.

For a $\mathcal{G}$-measurable random vector $\boldsymbol{z}$, the information in $\mathcal{G}$ does not improve on our understanding of $\boldsymbol{z}$, so that $\mathbb{E}(\boldsymbol{z} \mid \mathcal{G}) = \boldsymbol{z}$ with probability one. That is, $\boldsymbol{z}$ can be treated as known in $\mathbb{E}(\boldsymbol{z} \mid \mathcal{G})$ and taken out from the conditional expectation. Thus,

$$\mathbb{E}(\boldsymbol{z}\boldsymbol{x}' \mid \mathcal{G}) = \boldsymbol{z}\,\mathbb{E}(\boldsymbol{x}' \mid \mathcal{G}).$$

In particular, $\boldsymbol{z}$ can be taken out from the conditional expectation when $\boldsymbol{z}$ itself is a conditioning variable. This result is generalized as follows.

Lemma 5.10 *Let $\boldsymbol{z}$ be a $\mathcal{G}$-measurable random vector. Then for any Borel-measurable function g,*

$$\mathbb{E}[g(\boldsymbol{z})\boldsymbol{x} \mid \mathcal{G}] = g(\boldsymbol{z})\,\mathbb{E}(\boldsymbol{x} \mid \mathcal{G}),$$

with probability one.

Two square integrable random variables z and y are said to be *orthogonal* if their inner product $\mathbb{E}(zy) = 0$. This definition allows us to discuss orthogonal projection in the space of square integrable random vectors. Let z be a square integrable random variable and $\tilde{z}$ be a $\mathcal{G}$-measurable random variable. Then, by Lemma 5.9 (law of iterated expectations) and Lemma 5.10,

$$\begin{aligned}
\mathbb{E}\big[\big(z - \mathbb{E}(z \mid \mathcal{G})\big)\tilde{z}\big] &= \mathbb{E}\Big[\mathbb{E}\big[\big(z - \mathbb{E}(z \mid \mathcal{G})\big)\tilde{z} \mid \mathcal{G}\big]\Big] \\
&= \mathbb{E}\big[\mathbb{E}(z \mid \mathcal{G})\tilde{z} - \mathbb{E}(z \mid \mathcal{G})\tilde{z}\big] \\
&= 0.
\end{aligned}$$

That is, the difference between z and its conditional expectation $\mathbb{E}(z \mid \mathcal{G})$ must be orthogonal to any $\mathcal{G}$-measurable random variable. It can then be seen that for any square integrable, $\mathcal{G}$-measurable random variable $\tilde{z}$,

$$\begin{aligned}\mathbb{E}(z-\tilde{z})^2 &= \mathbb{E}[z-\mathbb{E}(z \mid \mathcal{G})+\mathbb{E}(z \mid \mathcal{G})-\tilde{z}]^2 \\ &= \mathbb{E}[z-\mathbb{E}(z \mid \mathcal{G})]^2+\mathbb{E}[\mathbb{E}(z \mid \mathcal{G})-\tilde{z}]^2 \\ &\geq \mathbb{E}[z-\mathbb{E}(z \mid \mathcal{G})]^2.\end{aligned}$$

where in the second equality the cross-product term vanishes because both $\mathbb{E}(z \mid \mathcal{G})$ and $\tilde{z}$ are $\mathcal{G}$-measurable and hence orthogonal to $z-\mathbb{E}(z \mid \mathcal{G})$. That is, among all $\mathcal{G}$-measurable random variables that are also square integrable, $\mathbb{E}(z \mid \mathcal{G})$ is the closest to z in terms of the L_2-norm. This shows that $\mathbb{E}(z \mid \mathcal{G})$ is the orthogonal projection of z onto the space of all $\mathcal{G}$-measurable, square integrable random variables.

Lemma 5.11 *Let z be a square integrable random variable. Then*

$$\mathbb{E}[z-\mathbb{E}(z \mid \mathcal{G})]^2 \leq \mathbb{E}(z-\tilde{z})^2,$$

for any $\mathcal{G}$-measurable random variable $\tilde{z}$.

In particular, let $\mathcal{G}=\sigma(\boldsymbol{y})$, where $\boldsymbol{y}$ is a square integrable random vector. Lemma 5.11 implies that

$$\mathbb{E}\big[z-\mathbb{E}\big(z \mid \sigma(\boldsymbol{y})\big)\big]^2 \leq \mathbb{E}\big(z-h(\boldsymbol{y})\big)^2,$$

for any Borel-measurable function h such that $h(\boldsymbol{y})$ is also square integrable. Thus, $\mathbb{E}[z \mid \sigma(\boldsymbol{y})]$ minimizes the L_2-norm $\|z-h(\boldsymbol{y})\|_2$, and its difference from z is orthogonal to any function of $\boldsymbol{y}$ that is also square integrable. We may then say that, given all the information generated from $\boldsymbol{y}$, $\mathbb{E}(z \mid \sigma(\boldsymbol{y}))$ is the "best approximation" of z in terms of the L_2-norm (the best L_2 predictor).

The *conditional variance-covariance matrix* of $\boldsymbol{z}$ given $\boldsymbol{y}$ is

$$\begin{aligned}\operatorname{var}(\boldsymbol{z} \mid \boldsymbol{y}) &= \mathbb{E}\big([\boldsymbol{z}-\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y})][\boldsymbol{z}-\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y})]' \mid \boldsymbol{y}\big) \\ &= \mathbb{E}(\boldsymbol{z}\boldsymbol{z}' \mid \boldsymbol{y})-\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y})\,\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y})'.\end{aligned}$$

Similar to unconditional variance-covariance matrix, we have for non-stochastic matrices $\boldsymbol{A}$ and $\boldsymbol{b}$,

$$\operatorname{var}(\boldsymbol{A}\boldsymbol{z}+\boldsymbol{b} \mid \boldsymbol{y}) = \boldsymbol{A}\,\operatorname{var}(\boldsymbol{z} \mid \boldsymbol{y})\,\boldsymbol{A}',$$

which is nonsingular provided that $\boldsymbol{A}$ has full row rank and $\mathrm{var}(\boldsymbol{z} \mid \boldsymbol{y})$ is positive definite. It can also be shown that

$$\mathrm{var}(\boldsymbol{z}) = \mathbb{E}[\mathrm{var}(\boldsymbol{z} \mid \boldsymbol{y})] + \mathrm{var}\big(\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y})\big);$$

see Exercise 5.7. That is, the variance of $\boldsymbol{y}$ can be expressed as the sum of two components: the mean of its conditional variance and the variance of its conditional mean. This is also known as the decomposition of *analysis of variance*.

Example 5.12 Suppose that $(\boldsymbol{y}' \ \boldsymbol{x}')'$ is distributed as a multivariate normal random vector:

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu_y} \\ \boldsymbol{\mu_x} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma_y} & \boldsymbol{\Sigma}'_{\boldsymbol{xy}} \\ \boldsymbol{\Sigma_{xy}} & \boldsymbol{\Sigma_x} \end{bmatrix} \right).$$

It is well known that the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$ is also normal. Moreover, it can be shown that

$$\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{x}) = \boldsymbol{\mu_y} - \boldsymbol{\Sigma}'_{\boldsymbol{xy}}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}(\boldsymbol{x} - \boldsymbol{\mu_x}),$$

a linear function of $\boldsymbol{x}$. By the analysis-of-variance decomposition, the conditional variance-covariance matrix of $\boldsymbol{y}$ is

$$\mathrm{var}(\boldsymbol{y} \mid \boldsymbol{x}) = \mathrm{var}(\boldsymbol{y}) - \mathrm{var}\big(\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{x})\big) = \boldsymbol{\Sigma_y} - \boldsymbol{\Sigma}'_{\boldsymbol{xy}}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma_{xy}},$$

which does not depend on $\boldsymbol{x}$.

5.3 Modes of Convergence

Consider now a sequence of random variables $\{z_n(\omega)\}_{n=1,2,\ldots}$ defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For a given ω, $\{z_n\}$ is a realization (a sequence of sample values) of the random element ω with the index n, and that for a given n, z_n is a random variable which assumes different values depending on ω. In this section we will discuss various modes of convergence for sequences of random variables.

5.3.1 Almost Sure Convergence

We first introduce the concept of *almost sure convergence* (*convergence with probability one*). Suppose that $\{z_n\}$ is a sequence of random variables and z is a random variable, all defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The sequence $\{z_n\}$ is said to converge to z almost surely if, and only if,

$$\mathbb{P}(\omega \colon z_n(\omega) \to z(\omega) \text{ as } n \to \infty) = 1,$$

denoted as $z_n \xrightarrow{\text{a.s.}} z$ or $z_n \to z$ a.s. Note that for a given ω, the realization $z_n(\omega)$ may or may not converge to $z(\omega)$. Almost sure convergence requires that $z_n(\omega) \to z(\omega)$ for almost all $\omega \in \Omega$, except for those ω in a set with probability zero. That is, almost all the realizations $z_n(\omega)$ will be eventually close to $z(\omega)$ for all n sufficiently large; the event that z_n will not approach z is improbable. When $\boldsymbol{z}_n$ and $\boldsymbol{z}$ are both $\mathbb{R}^d$-valued, almost sure convergence is defined elementwise. That is, $\boldsymbol{z}_n \to \boldsymbol{z}$ a.s. if every element of $\boldsymbol{z}_n$ converges almost surely to the corresponding element of $\boldsymbol{z}$.

The following result shows that continuous transformation preserves almost sure convergence.

Lemma 5.13 *Let $g\colon \mathbb{R} \mapsto \mathbb{R}$ be a function continuous on $S_g \subseteq \mathbb{R}$.*

[a] *If $z_n \xrightarrow{\text{a.s.}} z$, where z is a random variable such that $\mathbb{P}(z \in S_g) = 1$, then $g(z_n) \xrightarrow{\text{a.s.}} g(z)$.*

[b] *If $z_n \xrightarrow{\text{a.s.}} c$, where c is a real number at which g is continuous, then $g(z_n) \xrightarrow{\text{a.s.}} g(c)$.*

Proof: Let $\Omega_0 = \{\omega\colon z_n(\omega) \to z(\omega)\}$ and $\Omega_1 = \{\omega\colon z(\omega) \in S_g\}$. Thus, for $\omega \in (\Omega_0 \cap \Omega_1)$, continuity of g ensures that $g(z_n(\omega)) \to g(z(\omega))$. Note that

$$(\Omega_0 \cap \Omega_1)^c = \Omega_0^c \cup \Omega_1^c,$$

which has probability zero because $\mathbb{P}(\Omega_0^c) = \mathbb{P}(\Omega_1^c) = 0$. It follows that $\Omega_0 \cap \Omega_1$ has probability one. This proves that $g(z_n) \to g(z)$ with probability one. The second assertion is just a special case of the first result. □

Lemma 5.13 is easily generalized to $\mathbb{R}^d$-valued random variables. For example, $\boldsymbol{z}_n \xrightarrow{\text{a.s.}} \boldsymbol{z}$ implies

$$\begin{aligned} z_{1,n} + z_{2,n} &\xrightarrow{\text{a.s.}} z_1 + z_2, \\ z_{1,n} z_{2,n} &\xrightarrow{\text{a.s.}} z_1 z_2, \\ z_{1,n}^2 + z_{2,n}^2 &\xrightarrow{\text{a.s.}} z_1^2 + z_2^2, \end{aligned}$$

where $z_{1,n}, z_{2,n}$ are two elements of $\boldsymbol{z}_n$ and z_1, z_2 are the corresponding elements of $\boldsymbol{z}$. Also, provided that $z_2 \neq 0$ with probability one, $z_{1,n}/z_{2,n} \to z_1/z_2$ a.s.

5.3.2 Convergence in Probability

A weaker convergence concept is *convergence in probability*. A sequence of random variables $\{z_n\}$ is said to converge to z in probability if for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\omega\colon |z_n(\omega) - z(\omega)| > \epsilon) = 0,$$

or equivalently,

$$\lim_{n\to\infty} \mathbb{P}(\omega\colon |z_n(\omega) - z(\omega)| \le \epsilon) = 1,$$

denoted as $z_n \xrightarrow{\mathbb{P}} z$. We also say that z is the *probability limit* of z_n, denoted as $\operatorname{plim} z_n = z$. In particular, if the probability limit of z_n is a constant c, all the probability mass of z_n will concentrate around c when n becomes large. For $\mathbb{R}^d$-valued random variables $\boldsymbol{z}_n$ and $\boldsymbol{z}$, convergence in probability is also defined elementwise.

In the definition of convergence in probability, the events $\Omega_n(\epsilon) = \{\omega : |z_n(\omega) - z(\omega)| \le \epsilon\}$ vary with n, and convergence is referred to the probabilities of such events: $p_n = \mathbb{P}(\Omega_n(\epsilon))$, rather than the random variables z_n. By contrast, almost sure convergence is related directly to the behaviors of random variables. For convergence in probability, the event Ω_n that z_n will be close to z becomes highly likely when n tends to infinity, or its complement (z_n will deviate from z by a certain distance) becomes highly unlikely when n tends to infinity. Whether z_n will converge to z is not of any concern in convergence in probability.

More specifically, let Ω_0 denote the set of ω such that $z_n(\omega)$ converges to $z(\omega)$. For $\omega \in \Omega_0$, there is some m such that ω is in $\Omega_n(\epsilon)$ for all $n > m$. That is,

$$\Omega_0 \subseteq \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \Omega_n(\epsilon) \in \mathcal{F}.$$

As $\cap_{n=m}^{\infty}\Omega_n(\epsilon)$ is also in $\mathcal{F}$ and non-decreasing in m, it follows that

$$\mathbb{P}(\Omega_0) \le \mathbb{P}\left(\bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} \Omega_n(\epsilon)\right) = \lim_{m\to\infty} \mathbb{P}\left(\bigcap_{n=m}^{\infty} \Omega_n(\epsilon)\right) \le \lim_{m\to\infty} \mathbb{P}\big(\Omega_m(\epsilon)\big).$$

This inequality proves that almost sure convergence implies convergence in probability, but the converse is not true in general. We state this result below.

Lemma 5.14 *If* $z_n \xrightarrow{\text{a.s.}} z$, *then* $z_n \xrightarrow{\mathbb{P}} z$.

The following well-known example shows that when there is convergence in probability, the random variables themselves may not even converge for any ω.

Example 5.15 Let $\Omega = [0,1]$ and $\mathbb{P}$ be the Lebesgue measure (i.e., $\mathbb{P}\{(a,b]\} = b - a$ for $(a,b] \subseteq [0,1]$). Consider the sequence $\{I_n\}$ of intervals $[0,1]$, $[0,1/2)$, $[1/2,1]$, $[0,1/3)$, $[1/3,2/3)$, $[2/3,1]$, ... , and let $z_n = \mathbf{1}_{I_n}$ be the indicator function of I_n: $z_n(\omega) = 1$ if $\omega \in I_n$ and $z_n = 0$ otherwise. When n tends to infinity, I_n shrinks toward a singleton which has the Lebesgue measure zero. For $0 < \epsilon < 1$, we then have

$$\mathbb{P}(|z_n| > \epsilon) = \mathbb{P}(I_n) \to 0,$$

which shows $z_n \xrightarrow{\mathbb{P}} 0$. On the other hand, it is easy to see that each $\omega \in [0,1]$ must be covered by infinitely many intervals. Thus, given any $\omega \in [0,1]$, $z_n(\omega) = 1$ for infinitely many n, and hence $z_n(\omega)$ does not converge to zero. Note that convergence in probability permits z_n to deviate from the probability limit infinitely often, but almost sure convergence does not, except for those ω in the set of probability zero. □

Intuitively, if $\mathrm{var}(z_n)$ vanishes asymptotically, the distribution of z_n would shrink toward its mean $\mathbb{E}(z_n)$. If, in addition, $\mathbb{E}(z_n)$ tends to a constant c (or $\mathbb{E}(z_n) = c$), then z_n ought to be degenerate at c in the limit. These observations suggest the following sufficient conditions for convergence in probability; see Exercises 5.8 and 5.9. In many cases, it is easier to establish convergence in probability by verifying these conditions.

Lemma 5.16 *Let $\{z_n\}$ be a sequence of square integrable random variables. If $\mathbb{E}(z_n) \to c$ and $\mathrm{var}(z_n) \to 0$, then $z_n \xrightarrow{\mathbb{P}} c$.*

Analogous to Lemma 5.13, continuous functions also preserve convergence in probability.

Lemma 5.17 *Let $g\colon \mathbb{R}t\ \mathbb{R}$ be a function continuous on $S_g \subseteq \mathbb{R}$.*

[a] *If $z_n \xrightarrow{\mathbb{P}} z$, where z is a random variable such that $\mathbb{P}(z \in S_g) = 1$, then $g(z_n) \xrightarrow{\mathbb{P}} g(z)$.*

[b] **(Slutsky)** *If $z_n \xrightarrow{\mathbb{P}} c$, where c is a real number at which g is continuous, then $g(z_n) \xrightarrow{\mathbb{P}} g(c)$.*

Proof: By the continuity of g, for each $\epsilon > 0$, we can find a $\delta > 0$ such that

$$\{\omega : |z_n(\omega) - z(\omega)| \le \delta\} \cap \{\omega : z(\omega) \in S_g\}$$
$$\subseteq \{\omega : |g(z_n(\omega)) - g(z(\omega))| \le \epsilon\}.$$

Taking complementation of both sides and noting that the complement of $\{\omega : z(\omega) \in S_g\}$ has probability zero, we have

$$\mathbb{P}(|g(z_n) - g(z)| > \epsilon) \leq \mathbb{P}(|z_n - z| > \delta).$$

As z_n converges to z in probability, the right-hand side converges to zero and so does the left-hand side. □

Lemma 5.17 is readily generalized to $\mathbb{R}^d$-valued random variables. For instance, $\boldsymbol{z}_n \xrightarrow{\mathbb{P}} \boldsymbol{z}$ implies

$$\begin{aligned} z_{1,n} + z_{2,n} &\xrightarrow{\mathbb{P}} z_1 + z_2, \\ z_{1,n} z_{2,n} &\xrightarrow{\mathbb{P}} z_1 z_2, \\ z_{1,n}^2 + z_{2,n}^2 &\xrightarrow{\mathbb{P}} z_1^2 + z_2^2, \end{aligned}$$

where $z_{1,n}, z_{2,n}$ are two elements of $\boldsymbol{z}_n$ and z_1, z_2 are the corresponding elements of $\boldsymbol{z}$. Also, provided that $z_2 \neq 0$ with probability one, $z_{1,n}/z_{2,n} \xrightarrow{\mathbb{P}} z_1/z_2$.

5.3.3 Convergence in Distribution

Another convergence mode, known as *convergence in distribution* or *convergence in law*, concerns the behavior of the distribution functions of random variables. Let F_{z_n} and F_z be the distribution functions of z_n and z, respectively. A sequence of random variables $\{z_n\}$ is said to converge to z in distribution, denoted as $z_n \xrightarrow{D} z$, if

$$\lim_{n\to\infty} F_{z_n}(\zeta) = F_z(\zeta),$$

for every continuity point ζ of F_z. That is, regardless the distributions of z_n, convergence in distribution ensures that F_{z_n} will be arbitrarily close to F_z for all n sufficiently large. The distribution F_z is thus known as the *limiting distribution* of z_n. We also say that z_n is asymptotically distributed as F_z, denoted as $z_n \overset{A}{\sim} F_z$.

For random vectors $\{\boldsymbol{z}_n\}$ and $\boldsymbol{z}$, $\boldsymbol{z}_n \xrightarrow{D} \boldsymbol{z}$ if the joint distributions $F_{\boldsymbol{z}_n}$ converge to $F_{\boldsymbol{z}}$ for every continuity point $\boldsymbol{\zeta}$ of $F_{\boldsymbol{z}}$. It is, however, more cumbersome to show convergence in distribution for a sequence of random vectors. The so-called *Cramér-Wold device* allows us to transform this multivariate convergence problem to a univariate one. This result is stated below without proof.

Lemma 5.18 (Cramér-Wold Device) *Let $\{\boldsymbol{z}_n\}$ be a sequence of random vectors in $\mathbb{R}^d$. Then $\boldsymbol{z}_n \xrightarrow{D} \boldsymbol{z}$ if and only if $\boldsymbol{\alpha}'\boldsymbol{z}_n \xrightarrow{D} \boldsymbol{\alpha}'\boldsymbol{z}$ for every $\boldsymbol{\alpha} \in \mathbb{R}^d$ such that $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$.*

There is also a uni-directional relationship between convergence in probability and convergence in distribution. To see this, note that for some arbitrary $\epsilon > 0$ and a continuity point ζ of F_z, we have

$$\begin{aligned}\mathbb{P}(z_n \le \zeta) &= \mathbb{P}(\{z_n \le \zeta\} \cap \{|z_n - z| \le \epsilon\}) + \mathbb{P}(\{z_n \le \zeta\} \cap \{|z_n - z| > \epsilon\}) \\ &\le \mathbb{P}(z \le \zeta + \epsilon) + \mathbb{P}(|z_n - z| > \epsilon).\end{aligned}$$

Similarly,

$$\mathbb{P}(z \le \zeta - \epsilon) \le \mathbb{P}(z_n \le \zeta) + \mathbb{P}(|z_n - z| > \epsilon).$$

If $z_n \xrightarrow{\mathbb{P}} z$, then by passing to the limit and noting that ϵ is arbitrary, the inequalities above imply

$$\lim_{n\to\infty} \mathbb{P}(z_n \le \zeta) = \mathbb{P}(z \le \zeta).$$

That is, $F_{z_n}(\zeta) \to F_z(\zeta)$. The converse is not true in general, however.

When z_n converges in distribution to a real number c, it is not difficult to show that z_n also converges to c in probability. In this case, these two convergence modes are equivalent. To be sure, note that a real number c can be viewed as a degenerate random variable with the distribution function:

$$F(\zeta) = \begin{cases} 0, & \zeta < c, \\ 1, & \zeta \ge c, \end{cases}$$

which is a step function with a jump point at c. When $z_n \xrightarrow{D} c$, all the probability mass of z_n will concentrate at c as n becomes large; this is precisely what $z_n \xrightarrow{\mathbb{P}} c$ means. More formally, for any $\epsilon > 0$,

$$\mathbb{P}(|z_n - c| > \epsilon) = 1 - [F_{z_n}(c + \epsilon) - F_{z_n}((c - \epsilon)^-)],$$

where $(c-\epsilon)^-$ denotes the point adjacent to and less than $c - \epsilon$. Now, $z_n \xrightarrow{D} c$ implies that $F_{z_n}(c + \epsilon) - F_{z_n}((c - \epsilon)^-)$ converges to one, so that $\mathbb{P}(|z_n - c| > \epsilon)$ converges to zero. We summarizes these results below.

Lemma 5.19 *If $z_n \xrightarrow{\mathbb{P}} z$, then$z_n \xrightarrow{D} z$. For a constant c, $z_n \xrightarrow{\mathbb{P}} c$ is equivalent to $z_n \xrightarrow{D} c$.*

The *continuous mapping theorem* below asserts that continuous functions preserve convergence in distribution; cf. Lemmas 5.13 and 5.17.

Lemma 5.20 (Continuous Mapping Theorem) *Let $g\colon \mathbb{R} \to \mathbb{R}$ be a function continuous almost everywhere on $\mathbb{R}$, except for at most countably many points. If $z_n \xrightarrow{D} z$, then $g(z_n) \xrightarrow{D} g(z)$.*

For example, if z_n converges in distribution to the standard normal random variable, the limiting distribution of z_n^2 is $\chi^2(1)$. Generalizing this result to $\mathbb{R}^d$-valued random variables, we can see that when $\boldsymbol{z}_n$ converges in distribution to the d-dimensional standard normal random variable, the limiting distribution of $\boldsymbol{z}_n'\boldsymbol{z}_n$ is $\chi^2(d)$.

Two sequences of random variables $\{y_n\}$ and $\{z_n\}$ are said to be *asymptotically equivalent* if their differences $y_n - z_n$ converge to zero in probability. Intuitively, the limiting distributions of two asymptotically equivalent sequences, if exist, ought to be the same. This is stated in the next result without proof.

Lemma 5.21 *Let $\{y_n\}$ and $\{z_n\}$ be two sequences of random vectors such that $y_n - z_n \xrightarrow{\mathbb{P}} 0$. If $z_n \xrightarrow{D} z$, then $y_n \xrightarrow{D} z$.*

The next result is concerned with two sequences of random variables such that one converges in distribution and the other converges in probability.

Lemma 5.22 *If y_n converges in probability to a constant c and z_n converges in distribution to z, then $y_n + z_n \xrightarrow{D} c + z$, $y_n z_n \xrightarrow{D} cz$, and $z_n/y_n \xrightarrow{D} z/c$ if $c \neq 0$.*

5.4 Order Notations

It is typical to use *order notations* to describe the behavior of a sequence of numbers, whether it converges or not. Let $\{c_n\}$ denote a sequence of positive real numbers.

1. Given a sequence $\{b_n\}$, we say that b_n is (at most) of order c_n, denoted as $b_n = O(c_n)$, if there exists a $\Delta < \infty$ such that $|b_n|/c_n \leq \Delta$ for all sufficiently large n. When c_n diverges, b_n cannot diverge faster than c_n; when c_n converges to zero, the rate of convergence of b_n is no slower than that of c_n. For example, the polynomial $a + bn$ is $O(n)$, and the partial sum of a bounded sequence $\sum_{i=1}^n b_i$ is $O(n)$. Note that an $O(1)$ sequence is a bounded sequence.

2. Given a sequence $\{b_n\}$, we say that b_n is of smaller order than c_n, denoted as $b_n = o(c_n)$, if $b_n/c_n \to 0$. When c_n diverges, b_n must diverge slower than c_n; when c_n converges to zero, the rate of convergence of b_n should be faster than that of c_n.

For example, the polynomial $a + bn$ is $o(n^{1+\delta})$ for any $\delta > 0$, and the partial sum $\sum_{i=1}^{n} \alpha^i$, $|\alpha| < 1$, is $o(n)$. Note that an $o(1)$ sequence is a sequence that converges to zero.

If $\boldsymbol{b}_n$ is a vector (matrix), $\boldsymbol{b}_n$ is said to be $O(c_n)$ ($o(c_n)$) if every element of $\boldsymbol{b}_n$ is $O(c_n)$ ($o(c_n)$). It is also easy to verify the following results; see Exercise 5.10.

Lemma 5.23 *Let $\{a_n\}$ and $\{b_n\}$ be two non-stochastic sequences.*

(a) *If $a_n = O(n^r)$ and $b_n = O(n^s)$, then $a_n b_n = O(n^{r+s})$ and $a_n + b_n = O(n^{\max(r,s)})$.*

(b) *If $a_n = o(n^r)$ and $b_n = o(n^s)$, then $a_n b_n = o(n^{r+s})$ and $a_n + b_n = o(n^{\max(r,s)})$.*

(c) *If $a_n = O(n^r)$ and $b_n = o(n^s)$, then $a_n b_n = o(n^{r+s})$ and $a_n + b_n = O(n^{\max(r,s)})$.*

The order notations can be easily extended to describe the behavior of sequences of random variables. A sequence of random variables $\{z_n\}$ is said to be $O_{\text{a.s.}}(c_n)$ (or $O(c_n)$ almost surely) if z_n/c_n is $O(1)$ a.s., and it is said to be $O_{\mathbb{P}}(c_n)$ (or $O(c_n)$ in probability) if for every $\epsilon > 0$, there is some Δ such that

$$\mathbb{P}(|z_n|/c_n \geq \Delta) \leq \epsilon,$$

for all n sufficiently large. Similarly, $\{z_n\}$ is $o_{\text{a.s.}}(c_n)$ (or $o(c_n)$ almost surely) if $z_n/c_n \xrightarrow{\text{a.s.}} 0$, and it is $o_{\mathbb{P}}(c_n)$ (or $o(c_n)$ in probability) if $z_n/c_n \xrightarrow{\mathbb{P}} 0$.

If $\{z_n\}$ is $O_{\text{a.s.}}(1)$ ($o_{\text{a.s}}(1)$), we say that z_n is bounded (vanishing) almost surely; if $\{z_n\}$ is $O_{\mathbb{P}}(1)$ ($o_{\mathbb{P}}(1)$), z_n is bounded (vanishing) in probability. Note that Lemma 5.23 also holds for stochastic order notations. In particular, if a sequence of random variables is bounded almost surely (in probability) and another sequence of random variables is vanishing almost surely (in probability), the products of their corresponding elements are vanishing almost surely (in probability). That is, $y_n = O_{\text{a.s.}}(1)$ and $z_n = o_{\text{a.s}}(1)$, then $y_n z_n$ is $o_{\text{a.s}}(1)$.

When $z_n \xrightarrow{D} z$, we have known that z_n does not converge in probability to z in general, but more can be said about the behavior of z_n. Let ζ be a continuity point of F_z. Then for any $\epsilon > 0$, we can choose a sufficiently large ζ such that $\mathbb{P}(|z| > \zeta) < \epsilon/2$. As $z_n \xrightarrow{D} z$, we can also choose n large enough such that

$$\mathbb{P}(|z_n| > \zeta) - \mathbb{P}(|z| > \zeta) < \epsilon/2,$$

which implies $\mathbb{P}(|z_n| > \zeta) < \epsilon$. This shows that:

Lemma 5.24 *Let $\{z_n\}$ be a sequence of random vectors such that $z_n \xrightarrow{D} z$. Then $z_n = O_{\mathbb{P}}(1)$.*

5.5 Law of Large Numbers

The law of large numbers is concerned with the averaging behavior of random variables and is one of the most important results in probability theory. A sequence of integrable random variables $\{z_t\}$ is said to obey a *strong law of large numbers* (SLLN) if

$$\frac{1}{T}\sum_{t=1}^{T}[z_t - \mathbb{E}(z_t)] \xrightarrow{\text{a.s.}} 0, \tag{5.1}$$

and it is said to obey a *weak law of large numbers* (WLLN) if

$$\frac{1}{T}\sum_{t=1}^{T}[z_t - \mathbb{E}(z_t)] \xrightarrow{\mathbb{P}} 0. \tag{5.2}$$

Thus, a law of large numbers ensures that the sample average essentially follows its mean behavior; random irregularities (deviations from the mean) are eventually "wiped out" by averaging. For a sequence of random vectors (matrices), a SLLN (WLLN) is defined elementwise.

According to these definitions, z_t may or may not be independent random variables, and they are not required to have a constant mean and hence may have non-identical distributions. When $\mathbb{E}(z_t)$ is a constant μ_o, the SLLN (5.1) and WLLN (5.2) read, respectively,

$$\frac{1}{T}\sum_{t=1}^{T} z_t \xrightarrow{\text{a.s.}} \mu_o, \qquad \frac{1}{T}\sum_{t=1}^{T} z_t \xrightarrow{\mathbb{P}} \mu_o.$$

Below are two well known strong laws.

Lemma 5.25 (Kolmogorov) *Let $\{z_t\}$ be a sequence of i.i.d. random variables with mean μ_o. Then*

$$\frac{1}{T}\sum_{t=1}^{T} z_t \xrightarrow{\text{a.s.}} \mu_o.$$

Note that i.i.d. random variables need not obey Kolmogorov's SLLN if they do not have a finite mean, e.g., the variables with the Cauchy distribution. Given that almost sure convergence implies convergence in probability, the same condition in Lemma 5.25 ensures that $\{z_t\}$ obeys a WLLN.

Lemma 5.26 (Markov) *Let $\{z_t\}$ be a sequence of independent random variables. If for some $\delta > 0$, $\mathbb{E}\,|z_t|^{1+\delta}$ are bounded for all t, then*

$$\frac{1}{T}\sum_{t=1}^{T}[z_t - \mathbb{E}(z_t)] \xrightarrow{\text{a.s.}} 0,$$

From this result we can see that independent random variables may still obey a SLLN even when they do not have a common distribution. Comparing to Kolmogorov's SLLN, Lemma 5.26 requires random variables to satisfy a stronger moment condition (their $(1+\delta)^{\text{th}}$ moment must be bounded). A non-stochastic sequence, which can be viewed as a sequence of independent random variables, obeys a SLLN if it is $O(1)$.

The results above show that a SLLN (WLLN) holds provided that random variables satisfy certain conditions. The sufficient conditions ensuring a SLLN (WLLN) are usually imposed on the moments and dependence structure of random variables. Specifically, $\{z_t\}$ would obey a SLLN (WLLN) if z_t have bounded moments up to some order and are asymptotically independent in a proper sense. In some cases, it suffices to require $\operatorname{corr}(z_t, z_{t-j})$ converging to zero sufficiently fast as $j \to \infty$, as shown in the example below. Intuitively, random variables without some bounded moment may behave wildly such that their random irregularities cannot be completely averaged out. For random variables with strong correlations, the variation of their partial sums may grow too rapidly and cannot be eliminated by simple averaging. Thus, a sequence of random variables must be "well behaved" to ensure a SLLN (WLLN).

Example 5.27 Suppose that y_t is generated as a weakly stationary AR(1) process:

$$y_t = \alpha_o y_{t-1} + \epsilon_t, \qquad t = 1, 2, \ldots,$$

with $y_0 = 0$ and $|\alpha_o| < 1$, where ϵ_t are i.i.d. random variables with mean zero and variance σ^2. In view of Section 4.4, we have $\mathbb{E}(y_t) = 0$, $\operatorname{var}(y_t) = \sigma^2/(1-\alpha_o^2)$, and

$$\operatorname{cov}(y_t, y_{t-j}) = \alpha_o^j \frac{\sigma^2}{1-\alpha_o^2}.$$

These results ensure that $\mathbb{E}(T^{-1}\sum_{t=1}^{T} y_t) = 0$ and

$$\begin{aligned}
\operatorname{var}\left(\sum_{t=1}^{T} y_t\right) &= \sum_{t=1}^{T}\operatorname{var}(y_t) + 2\sum_{\tau=1}^{T-1}(T-\tau)\operatorname{cov}(y_t, y_{t-\tau}) \\
&\le \sum_{t=1}^{T}\operatorname{var}(y_t) + 2T\sum_{\tau=1}^{T-1}|\operatorname{cov}(y_t, y_{t-\tau})| \\
&= O(T).
\end{aligned}$$

The latter result shows that $\text{var}\left(T^{-1}\sum_{t=1}^{T} y_t\right) = O(T^{-1})$ which converges to zero as T approaches infinity. It follows from Lemma 5.16 that

$$\frac{1}{T}\sum_{t=1}^{T} y_t \xrightarrow{\mathbb{P}} 0.$$

This shows that $\{y_t\}$ obeys a WLLN. Note that in this case, y_t have a constant variance and $\text{cov}(y_t, y_{t-j})$ goes to zero exponentially fast as j tends to infinity. These two properties in effect ensure a WLLN. Similarly, it can be shown that

$$\frac{1}{T}\sum_{t=1}^{T} y_t^2 \xrightarrow{\mathbb{P}} \mathbb{E}(y_t^2) = \text{var}(y_t).$$

That is, $\{y_t^2\}$ also obeys a WLLN. These results are readily generalized to weakly stationary AR(p) processes. □

It is more cumbersome to establish a strong law for weakly stationary processes. The lemma below is convenient in practice; see Davidson (1994, p. 326) for a proof.

Lemma 5.28 *Let $y_t = \sum_{j=-\infty}^{\infty} \pi_j u_{t-j}$, where u_t are i.i.d. random variables with mean zero and variance σ^2. If π_j are absolutely summable, i.e., $\sum_{j=-\infty}^{\infty} |\pi_j| < \infty$, then $\sum_{t=1}^{T} y_t/T \xrightarrow{\text{a.s.}} 0$.*

In Example 5.27, $y_t = \sum_{j=0}^{\infty} \alpha_o^j \epsilon_{t-j}$ with $|\alpha_o| < 1$. It is clear that $\sum_{j=0}^{\infty} |\alpha_o^j| < \infty$. Hence, Lemma 5.28 ensures that $\{y_t\}$ obeys a SLLN and the average of y_t converges to its mean (zero) almost surely. If $y_t = z_t - \mu$, then the average of z_t converges to $\mathbb{E}(z_t) = \mu$ almost surely.

More generally, it is also possible that a sequence of weakly dependent and heterogeneously distributed random variables obeys a SLLN (WLLN). This usually requires stronger conditions on their moments and dependence structure.[1] To avoid technicality, we will not specify the regularity conditions that ensure a general SLLN (WLLN); see White (1984) and Davidson (1994) for such conditions and the resulting strong and weak laws. Instead, we use the following examples to illustrate why a WLLN and hence a SLLN may fail to hold.

[1]The notions of *mixing sequence* and *mixingale* allow the random variables to be dependent and heterogeneously distributed. In their definitions, probabilistic structures are imposed to regulate the dependence among random variables. Such sequences of random variables may obey a SLLN (WLLN) if they are *weakly dependent* in the sense that the dependence of random variables z_t on their distant past z_{t-j} eventually vanishes at a suitable rate as j tends to infinity.

Example 5.29 Consider the sequences $\{t\}$ and $\{t^2\}$, $t = 1, 2, \ldots$. It is well known that

$$\sum_{t=1}^{T} t = T(T+1)/2,$$
$$\sum_{t=1}^{T} t^2 = T(T+1)(2T+1)/6.$$

Hence, $\sum_{t=1}^{T} t/T$ and $\sum_{t=1}^{T} t^2/T$ both diverge. In this example, the elements of these two sequences diverge so that their partial sums grow too rapidly. Thus, these sequences do not obey a SLLN. □

Example 5.30 Suppose that ϵ_t are i.i.d. random variables with mean zero and variance σ^2. Thus, $T^{-1}\sum_{t=1}^{T} \epsilon_t \xrightarrow{\text{a.s.}} 0$ by Kolmogorv's SLLN (Lemma 5.25). As $\mathbb{E}\,|t\epsilon_t|^{1+\delta} = O(t^{1+\delta})$ which grows with t, $\{t\epsilon_t\}$ does not have bounded $(1+\delta)^{\text{th}}$ moment and therefore does not obey Markov's SLLN (Lemma 5.26). Moreover, note that

$$\operatorname{var}\left(\sum_{t=1}^{T} t\epsilon_t\right) = \sum_{t=1}^{T} t^2 \operatorname{var}(\epsilon_t) = \sigma^2 \frac{T(T+1)(2T+1)}{6}.$$

By Exercise 5.11, $\sum_{t=1}^{T} t\epsilon_t = O_{\mathbb{P}}(T^{3/2})$. It follows that $T^{-1}\sum_{t=1}^{T} t\epsilon_t = O_{\mathbb{P}}(T^{1/2})$ which diverges in probability. Thus, $\{t\epsilon_t\}$ does not obey a WLLN either. □

Example 5.31 Suppose that y_t is generated as a *random walk*:

$$y_t = y_{t-1} + \epsilon_t, \qquad t = 1, 2, \ldots,$$

with $y_0 = 0$, where ϵ_t are i.i.d. random variables with mean zero and variance σ^2. Clearly,

$$y_t = \sum_{i=1}^{t} \epsilon_i,$$

which has mean zero and unbounded variance $t\sigma^2$. For $s < t$, write

$$y_t = y_s + \sum_{i=s+1}^{t} \epsilon_i = y_s + v_{t-s},$$

where $v_{t-s} = \sum_{i=s+1}^{t} \epsilon_i$ is independent of y_s. We then have

$$\operatorname{cov}(y_t, y_s) = \mathbb{E}(y_s^2) = s\sigma^2,$$

for $t > s$. Consequently,

$$\operatorname{var}\left(\sum_{t=1}^{T} y_t\right) = \sum_{t=1}^{T} \operatorname{var}(y_t) + 2\sum_{\tau=1}^{T-1}\sum_{t=\tau+1}^{T} \operatorname{cov}(y_t, y_{t-\tau}).$$

It can be verified that the first term on the right-hand side is

$$\sum_{t=1}^{T} \operatorname{var}(y_t) = \sum_{t=1}^{T} t\sigma^2 = O(T^2),$$

and that the second term is

$$2\sum_{\tau=1}^{T-1}\sum_{t=\tau+1}^{T} \operatorname{cov}(y_t, y_{t-\tau}) = 2\sum_{\tau=1}^{T-1}\sum_{t=\tau+1}^{T} (t-\tau)\sigma^2 = O(T^3).$$

Thus, $\operatorname{var}(\sum_{t=1}^{T} y_t) = O(T^3)$, so that $\sum_{t=1}^{T} y_t = O_{\mathbb{P}}(T^{3/2})$ by Exercise 5.11. This shows that

$$\frac{1}{T}\sum_{t=1}^{T} y_t = O_{\mathbb{P}}(T^{1/2}),$$

which diverges in probability. Note that in this case, y_t have unbounded variances and strong correlations over time. Due to these correlations, the variation of the partial sum of y_t grows much too fast. (Recall that the variance of $\sum_{t=1}^{T} y_t$ is only $O(T)$ in Example 5.27.) Similarly, we can show that $\sum_{t=1}^{T} y_t^2 = O_{\mathbb{P}}(T^2)$; see Exercise 5.12 for a special case. Thus, $\{y_t^2\}$ does not obey a WLLN when y_t follows a random walk. As $\{y_t\}$ and $\{y_t^2\}$ do not obey a WLLN, they cannot obey a SLLN. The conclusions above will not be altered when $\{\epsilon_t\}$ is a white noise or a weakly stationary process. □

The example below shows that a sequence of random variables need not obey a WLLN even its partial sums are $O_{\mathbb{P}}(T)$.

Example 5.32 Suppose that y_t is generated as a random walk:

$$y_t = y_{t-1} + \epsilon_t, \qquad t = 1, 2, \ldots,$$

with $y_0 = 0$, as in Example 5.31. Then, the sequence $\{y_{t-1}\epsilon_t\}$ has mean zero and

$$\operatorname{var}(y_{t-1}\epsilon_t) = \mathbb{E}(y_{t-1}^2)\,\mathbb{E}(\epsilon_t^2) = (t-1)\sigma^4.$$

More interestingly, it can be seen that for $s < t$,

$$\operatorname{cov}(y_{t-1}\epsilon_t, y_{s-1}\epsilon_s) = \mathbb{E}(y_{t-1}y_{s-1}\epsilon_s)\,\mathbb{E}(\epsilon_t) = 0.$$

We then have

$$\text{var}\left(\sum_{t=1}^{T} y_{t-1}\epsilon_t\right) = \sum_{t=1}^{T} \text{var}(y_{t-1}\epsilon_t) = \sum_{t=1}^{T}(t-1)\sigma^4 = O(T^2),$$

and $\sum_{t=1}^{T} y_{t-1}\epsilon_t = O_{\mathbb{P}}(T)$. Note, however, that $\text{var}(T^{-1}\sum_{t=1}^{T} y_{t-1}\epsilon_t)$ converges to $\sigma^4/2$, rather than 0. Thus, $T^{-1}\sum_{t=1}^{T} y_{t-1}\epsilon_t$ cannot behave like a non-stochastic number in the limit. This shows that $\{y_{t-1}\epsilon_t\}$ does not obey a WLLN, and hence also does not obey a SLLN. □

5.6 Uniform Law of Large Numbers

In econometric analysis, it is also common to deal with functions of random variables and model parameters. For example, $q(z_t(\omega)); \theta)$ is a random variable for a given parameter θ, and it is function of θ for a given ω. When θ is fixed, it is not difficult to impose suitable conditions on q and z_t such that $\{q(z_t(\omega); \theta)\}$ obeys a SLLN (WLLN), as discussed in Section 5.5. When θ assumes values in the parameter space Θ, a SLLN (WLLN) that does not depend on θ is then needed.

More specifically, suppose that $\{q(z_t; \theta)\}$ obeys a SLLN for *each* $\theta \in \Theta$:

$$Q_T(\omega; \theta) = \frac{1}{T}\sum_{t=1}^{T} q(z_t(\omega); \theta) \xrightarrow{\text{a.s.}} Q(\theta),$$

where $Q(\theta)$ is a non-stochastic function of θ. As this convergent behavior may depend on θ, $\Omega_0^c(\theta) = \{\omega \colon Q_T(\omega; \theta) \not\to Q(\theta)\}$ varies with θ. When Θ is an interval of $\mathbb{R}$, $\cup_{\theta\in\Theta}\Omega_0^c(\theta)$ is an uncountable union of non-convergence sets and hence may not have probability zero, even though each $\Omega_0^c(\theta)$ does. Thus, the event that $Q_T(\omega; \theta) \to Q(\theta)$ for *all* θ, i.e., $\cap_{\theta\in\Theta}\Omega_0(\theta)$, may occur with probability less than one. In fact, the union of all $\Omega_0^c(\theta)$ may not even be in $\mathcal{F}$ (only countable unions of the elements in $\mathcal{F}$ are guaranteed to be in $\mathcal{F}$). If so, we cannot conclude anything about stochastic convergence. Worse still is when θ also depends on T, as in the case where θ is replaced by the estimator $\tilde{\theta}_T$. There may not exist a finite T^* such that $Q_T(\omega; \tilde{\theta}_T)$ are arbitrarily close to $Q(\omega; \tilde{\theta}_T)$ for all $T > T^*$.

These observations suggest that we should study convergence that is *uniform* on the parameter space Θ. In particular, $Q_T(\omega; \theta)$ converges to $Q(\theta)$ uniformly in θ almost surely (in probability) if the largest possible difference:

$$\sup_{\theta\in\Theta} |Q_T(\theta) - Q(\theta)| \to 0, \quad \text{a.s. (in probability)}.$$

In what follows we always assume that this supremum is a random variables for all T. The example below, similar to Example 2.14 of Davidson (1994), illustrates the difference between uniform and pointwise convergence.

Example 5.33 Let z_t be i.i.d. random variables with zero mean and

$$q_T(z_t(\omega);\theta) = z_t(\omega) + \begin{cases} T\theta, & 0 \le \theta \le \frac{1}{2T}, \\ 1 - T\theta, & \frac{1}{2T} < \theta \le \frac{1}{T}, \\ 0, & \frac{1}{T} < \theta < \infty. \end{cases}$$

Observe that for $\theta \ge 1/T$ and $\theta = 0$,

$$Q_T(\omega;\theta) = \frac{1}{T}\sum_{t=1}^{T} q_T(z_t;\theta) = \frac{1}{T}\sum_{t=1}^{T} z_t,$$

which converges to zero almost surely by Kolmogorov's SLLN. Thus, for a given θ, we can always choose T large enough such that $Q_T(\omega;\theta) \xrightarrow{\text{a.s.}} 0$, where 0 is the pointwise limit. On the other hand, it can be seen that $\Theta = [0,\infty)$ and

$$\sup_{\theta\in\Theta} |Q_T(\omega;\theta)| = |\bar{z}_T + 1/2| \xrightarrow{\text{a.s.}} 1/2,$$

so that the uniform limit is different from the pointwise limit. □

More generally, we consider a triangular array of functions $q_{Tt}(\boldsymbol{z}_t;\boldsymbol{\theta})$, $t = 1, 2, \ldots, T$, where $\boldsymbol{z}_t$ are integrable random vectors and $\boldsymbol{\theta}$ is the parameter vector taking values in the parameter space $\Theta \in \mathbb{R}^m$. For notation simplicity, we will not explicitly write ω in the functions. We say that $\{q_{Tt}(\boldsymbol{z}_t;\boldsymbol{\theta})\}$ obeys a *strong uniform law of large numbers* (SULLN) if

$$\sup_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T}[q_{Tt}(\boldsymbol{z}_t;\boldsymbol{\theta}) - \mathbb{E}(q_{Tt}(\boldsymbol{z}_t;\boldsymbol{\theta}))] \xrightarrow{\text{a.s.}} 0, \tag{5.3}$$

cf. (5.1). Similarly, $\{q_{Tt}(\boldsymbol{z}_t;\boldsymbol{\theta})\}$ is said to obey a *weak uniform law of large numbers* (WULLN) if the convergence condition above holds in probability. If q_{Tt} is $\mathbb{R}^m$-valued functions, the SULLN (WULLN) is defined elementwise.

We have seen that pointwise convergence does not imply uniform convergence. A natural question one would ask is: what additional conditions are needed to guarantee uniform convergence? Now let

$$Q_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}[q_{Tt}(\boldsymbol{z}_t;\boldsymbol{\theta}) - \mathbb{E}(q_{Tt}(\boldsymbol{z}_t;\boldsymbol{\theta}))].$$

Suppose that Q_T satisfies the following Lipschitz-type continuity requirement: for $\boldsymbol{\theta}, \boldsymbol{\theta}^\dagger$ in Θ,

$$|Q_T(\boldsymbol{\theta}) - Q_T(\boldsymbol{\theta}^\dagger)| \leq C_T \|\boldsymbol{\theta} - \boldsymbol{\theta}^\dagger\| \quad \text{a.s.},$$

where $\|\cdot\|$ denotes the Euclidean norm, and C_T is a random variable bounded almost surely and does not depend on $\boldsymbol{\theta}$. Under this condition, $Q_T(\boldsymbol{\theta}^\dagger)$ can be made arbitrarily close to $Q_T(\boldsymbol{\theta})$, provided that $\boldsymbol{\theta}^\dagger$ is sufficiently close to $\boldsymbol{\theta}$. Using the triangle inequality and taking supremum over $\boldsymbol{\theta}$ we have

$$\sup_{\boldsymbol{\theta}\in\Theta} |Q_T(\boldsymbol{\theta})| \leq \sup_{\boldsymbol{\theta}\in\Theta} |Q_T(\boldsymbol{\theta}) - Q_T(\boldsymbol{\theta}^\dagger)| + |Q_T(\boldsymbol{\theta}^\dagger)|.$$

Let Δ denote an almost sure bound of C_T. Then given any $\epsilon > 0$, choosing $\boldsymbol{\theta}^\dagger$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^\dagger\| < \epsilon/(2\Delta)$ implies

$$\sup_{\boldsymbol{\theta}\in\Theta} |Q_T(\boldsymbol{\theta}) - Q_T(\boldsymbol{\theta}^\dagger)| \leq C_T \frac{\epsilon}{2\Delta} \leq \frac{\epsilon}{2},$$

uniformly in T. Moreover, because $Q_T(\boldsymbol{\theta})$ converges to 0 almost surely for each $\boldsymbol{\theta}$ in Θ, $|Q_T(\boldsymbol{\theta}^\dagger)|$ is also less than $\epsilon/2$ for sufficiently large T. Consequently,

$$\sup_{\boldsymbol{\theta}\in\Theta} |Q_T(\boldsymbol{\theta})| \leq \epsilon,$$

for all T sufficiently large. As these results hold almost surely, we have a SULLN for $Q_T(\boldsymbol{\theta})$; the conditions ensuring a WULLN are analogous.

Lemma 5.34 *Suppose that for each $\boldsymbol{\theta} \in \Theta$, $\{q_{Tt}(\boldsymbol{z}_t; \boldsymbol{\theta})\}$ obeys a SLLN (WLLN) and that for $\boldsymbol{\theta}, \boldsymbol{\theta}^\dagger \in \Theta$,*

$$|Q_T(\boldsymbol{\theta}) - Q_T(\boldsymbol{\theta}^\dagger)| \leq C_T \|\boldsymbol{\theta} - \boldsymbol{\theta}^\dagger\| \quad \textit{a.s.},$$

where C_T is a random variable bounded almost surely (in probability) and does not depend on $\boldsymbol{\theta}$. Then, $\{q_{Tt}(\boldsymbol{z}_t; \boldsymbol{\theta})\}$ obeys a SULLN (WULLN).

Lemma 5.34 is quite convenient for establishing a SULLN (WULLN) because it requires only two conditions. First, the random functions must obey a standard SLLN (WLLN) for each $\boldsymbol{\theta}$ in the parameter space. Second, the random function must satisfy a Lipschitz-type continuity condition. In particular, the continuity condition can be verified by checking whether q_{Tt} is sufficiently "smooth" in the second argument. Note, however, that C_T being bounded almost surely may imply that the random variables in q_{Tt} are also bounded almost surely. This requirement is much too restrictive in

applications. Hence, a SULLN may not be readily obtained from Lemma 5.34. On the other hand, a WULLN is practically more plausible because the requirement that C_T is $O_{\mathbb{P}}(1)$ is much weaker. For example, the boundedness of $\mathbb{E}|C_T|$ is sufficient for C_T being $O_{\mathbb{P}}(1)$ by Markov's inequality. For more specific conditions ensuring these requirements we refer to Gallant and White (1988) and Bierens (1994).

5.7 Central Limit Theorem

The *central limit theorem* ensures that the distributions of suitably normalized averages will be essentially close to the standard normal distribution, regardless of the original distributions of random variables. This result is very useful and convenient in applications because, as far as approximation is concerned, we only have to consider a single distribution for normalized sample averages.

Given a sequence of square integrable random variables $\{z_t\}$, let $\bar{z}_T = T^{-1}\sum_{t=1}^T z_t$, $\bar{\mu}_T = T^{-1}\sum_{t=1}^T \mathbb{E}(z_t)$, and

$$\sigma_T^2 = \operatorname{var}\left(T^{-1/2}\sum_{t=1}^T z_t\right).$$

Then $\{z_t\}$ is said to obey a central limit theorem (CLT) if $\sigma_T^2 \to \sigma_o^2 > 0$ such that

$$\frac{1}{\sigma_o\sqrt{T}}\sum_{t=1}^T [z_t - \mathbb{E}(z_t)] = \frac{\sqrt{T}(\bar{z}_T - \bar{\mu}_T)}{\sigma_o} \xrightarrow{D} N(0,1). \tag{5.4}$$

Note that this definition requires neither $\mathbb{E}(z_t)$ nor $\operatorname{var}(z_t)$ to be a constant; also, $\{z_t\}$ may or may not be a sequence of independent variables. The following are two well known CLTs.

Lemma 5.35 (Lindeberg-Lévy) *Let $\{z_t\}$ be a sequence of i.i.d. random variables with mean μ_o and variance $\sigma_o^2 > 0$. Then,*

$$\frac{\sqrt{T}(\bar{z}_T - \mu_o)}{\sigma_o} \xrightarrow{D} N(0,1).$$

A sequence of i.i.d. random variables need not obey this CLT if they do not have a finite variance, e.g., random variables with $t(2)$ distribution. Comparing to Lemma 5.25, one can immediately see that the Lindeberg-Lévy CLT requires a stronger condition (i.e., finite variance) than does Kolmogorov's SLLN.

Remark: In this example, $\bar{z}_T$ converges to μ_o in probability, and its variance σ^2/T vanishes when T tends to infinity. To prevent having a degenerate distribution in the limit, it is then natural to consider the normalized average $T^{1/2}(\bar{z}_T - \mu_o)$, which has a constant variance σ^2 for all T. This explains why the normalizing factor $T^{1/2}$ is needed. For a normalizing factor T^a with $a < 1/2$, the normalized average still converges to zero because its variance vanishes in the limit. For a normalizing factor T^a with $a > 1/2$, the normalized average diverges. In both cases, the resulting normalized averages cannot have a well-behaved, non-degenerate distribution in the limit. Thus, it is usually said that $\bar{z}_T$ converges to μ_o at the rate $T^{-1/2}$.

Lemma 5.36 *Let $\{z_{Tt}\}$ be a triangular array of independent random variables with mean μ_{Tt} and variance $\sigma_{Tt}^2 > 0$ such that*

$$\bar{\sigma}_T^2 = \frac{1}{T}\sum_{t=1}^{T} \sigma_{Tt}^2 \to \sigma_o^2 > 0.$$

If for some $\delta > 0$, $\mathbb{E}\,|z_{Tt}|^{2+\delta}$ are bounded for all t, then

$$\frac{\sqrt{T}(\bar{z}_T - \bar{\mu}_T)}{\sigma_o} \xrightarrow{D} N(0,1).$$

Lemma 5.36 is a version of Liapunov's CLT. Note that this result requires a stronger condition (the $(2+\delta)^{\text{th}}$ moment must be bounded) than does Markov's SLLN (Lemma 5.26).

The sufficient conditions that ensure a CLT are similar to but usually stronger than those for a WLLN. That is, the sequence of random variables must have bounded moment up to some higher order, and random variables must be asymptotically independent of those in the distant past (such dependence must vanish sufficiently fast). Moreover, it is also required that every random variable in the sequence is asymptotically negligible, in the sense that no random variable is influential in affecting the partial sums. Although we will not specify these regularity conditions explicitly, we note that weakly stationary AR and MA processes obey a CLT in general. A sequence of weakly dependent and heterogeneously distributed random variables may also obey a CLT, depending on its moment and dependence structure. The following are examples that a CLT does not hold.

Example 5.37 Suppose that $\{\epsilon_t\}$ is a sequence of independent random variables with mean zero, variance σ^2, and bounded $(2+\delta)^{\text{th}}$ moment. From Example 5.29, we know

$\text{var}(\sum_{t=1}^T t\epsilon_t)$ is $O(T^3)$, which implies that $T^{-1/2}\sum_{t=1}^T t\epsilon_t$ still has a diverging variance (of order $O(T^2)$). On the other hand, observe that

$$\text{var}\left(\frac{1}{T^{1/2}}\sum_{t=1}^T \frac{t}{T}\epsilon_t\right) = \frac{T(T+1)(2T+1)}{6T^3}\sigma^2 \to \frac{\sigma^2}{3}.$$

It follows from Lemma 5.36 that

$$\frac{\sqrt{3}}{T^{1/2}\sigma}\sum_{t=1}^T \frac{t}{T}\epsilon_t \xrightarrow{D} N(0,1).$$

These results show that $\{(t/T)\epsilon_t\}$ obeys a CLT, whereas $\{t\epsilon_t\}$ does not. □

Example 5.38 Suppose that y_t is generated as a random walk:

$$y_t = y_{t-1} + \epsilon_t, \qquad t = 1, 2, \ldots,$$

with $y_0 = 0$, where $\{\epsilon_t\}$ is a sequence of i.i.d. random variables with mean zero and variance σ^2. From Example 5.31 we have seen that $\{y_t\}$ and $\{y_t^2\}$ have unbounded variances and strong correlations over time. Hence, they do not obey a CLT. Example 5.32 also suggests that $\{y_{t-1}\epsilon_t\}$ does not obey a CLT. □

Given a sequence of square integrable random vectors $\{\boldsymbol{z}_t\}$ in $\mathbb{R}^d$, let

$$\bar{\boldsymbol{z}}_T = \frac{1}{T}\sum_{t=1}^T \boldsymbol{z}_t,$$

$\bar{\boldsymbol{\mu}}_T = T^{-1}\sum_{t=1}^T \mathbb{E}(\boldsymbol{z}_t)$, and

$$\boldsymbol{\Sigma}_T = \text{var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T \boldsymbol{z}_t\right).$$

Then, $\{\boldsymbol{z}_t\}$ is said to obey a CLT if $\boldsymbol{\Sigma}_T \to \boldsymbol{\Sigma}_o$, a positive definite matrix, such that

$$\boldsymbol{\Sigma}_o^{-1/2}\frac{1}{\sqrt{T}}\sum_{t=1}^T[\boldsymbol{z}_t - \mathbb{E}(\boldsymbol{z}_t)] = \boldsymbol{\Sigma}_o^{-1/2}\sqrt{T}(\bar{\boldsymbol{z}}_T - \bar{\boldsymbol{\mu}}_T) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_d).$$

By Lemma 5.18, this amounts to requiring that $\{\boldsymbol{\alpha}'\boldsymbol{z}_t\}$ obeys a CLT, for any $\boldsymbol{\alpha} \in \mathbb{R}^d$ such that $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$.

Exercises

5.1 Let $\mathcal{C}$ be a collection of subsets of Ω. Show that the intersection of all the σ-algebras on Ω that contain $\mathcal{C}$ is the smallest σ-algebra containing $\mathcal{C}$.

5.2 Show that any half lines $(-\infty, b]$ and $[a, \infty)$ can be generated by open intervals in $\mathbb{R}$. Also show that any open interval (a, b) can be generated by closed intervals in $\mathbb{R}$.

5.3 Let y and z be two independent, integrable random variables. Show that $\mathbb{E}(yz) = \mathbb{E}(y)\,\mathbb{E}(z)$.

5.4 Let x and y be two random variables with finite p^{th} moment ($p > 1$). Prove the following triangle inequality:

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p.$$

Hint: Write $\mathbb{E}\,|x + y|^p = \mathbb{E}(|x + y||x + y|^{p-1})$ and apply Hölder's inequality.

5.5 In the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ suppose that we know the event B in $\mathcal{F}$ has occurred. Show that the conditional probability $\mathbb{P}(\cdot|B)$ satisfies the axioms for probability measures.

5.6 Prove the first assertion of Lemma 5.9.

5.7 Prove that for the square integrable random vectors $\boldsymbol{z}$ and $\boldsymbol{y}$,

$$\text{var}(\boldsymbol{z}) = \mathbb{E}[\text{var}(\boldsymbol{z} \mid \boldsymbol{y})] + \text{var}(\mathbb{E}(\boldsymbol{z} \mid \boldsymbol{y})).$$

5.8 A sequence of square integrable random variables $\{z_n\}$ is said to converge to a random variable z in L_2 (in quadratic mean) if

$$\mathbb{E}(z_n - z)^2 \to 0.$$

Prove that L_2 convergence implies convergence in probability.
Hint: Apply Chebychev's inequality.

5.9 Show that a sequence of square integrable random variables $\{z_n\}$ converges to a constant c in L_2 if and only if $\mathbb{E}(z_n) \to c$ and $\text{var}(z_n) \to 0$.

5.10 Prove Lemma 5.23.

5.11 Suppose that $\mathbb{E}(z_n^2) = O(c_n)$, where $\{c_n\}$ is a sequence of positive real numbers. Show that $z_n = O_{\mathbb{P}}(c_n^{1/2})$.

5.12 Suppose that y_t is generated as a Gaussian random walk:

$$y_t = y_{t-1} + \epsilon_t, \qquad t = 1, 2, \ldots ,$$

with $y_0 = 0$, where $\{\epsilon_t\}$ is a sequence of i.i.d. normal random variables with mean zero and variance σ^2. Show that $\sum_{t=1}^{T} y_t^2$ is $O_{\mathbb{P}}(T^2)$.

References

Bierens, Herman J. (1994). *Topics in Advanced Econometrics*, New York, NY: Cambridge University Press.

Davidson, James (1994). *Stochastic Limit Theory*, New York, NY: Oxford University Press.

Gallant, A. Ronald (1997). *An Introduction to Econometric Theory*, Princeton, NJ: Princeton University Press.

Gallant, A. Ronald and Halbert White (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*, Oxford, UK: Basil Blackwell.

White, Halbert (1984). *Asymptotic Theory for Econometricians*, Orlando, FL: Academic Press.

Chapter 6

Asymptotic Least Squares Theory

We have shown that the OLS estimator and related tests have good finite-sample properties under the classical conditions. These conditions are, however, quite restrictive in practice, as discussed in Section 3.7. It is therefore natural to ask the following questions. First, to what extent may we relax the classical conditions so that the OLS method has broader applicability? Second, what are the properties of the OLS method under more general conditions? The purpose of this chapter is to provide some answers to these questions. In particular, the analysis in this chapter allows the observations of each explanatory variable to be random variables, possibly weakly dependent and heterogeneously distributed. This relaxation permits applications of the OLS method to various data and models, but it also renders the analysis of finite-sample properties difficult. Nonetheless, it is relatively easy to analyze the asymptotic performance of the OLS estimator and construct large-sample tests. As the asymptotic results are valid under more general conditions, the OLS method remains a useful tool in a wide variety of applications.

6.1 When Regressors are Stochastic

Given the linear specification $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, suppose now that $\boldsymbol{X}$ is stochastic. In this case, [A2](i) must also be modified because $\mathbb{E}(\boldsymbol{y})$ cannot be a random vector $\boldsymbol{X}\boldsymbol{\beta}_o$. Even a condition on $\mathbb{E}(\boldsymbol{y})$ is available, we are still unable to evaluate

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_T) = \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}],$$

because $\hat{\boldsymbol{\beta}}_T$ now is a complex function of the elements of $\boldsymbol{y}$ and $\boldsymbol{X}$. Similarly, a condition on $\mathrm{var}(\boldsymbol{y})$ is of little use for calculating $\mathrm{var}(\hat{\boldsymbol{\beta}}_T)$.

To ensure unbiasedness, it is typical to impose the condition: $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}_o$ for some $\boldsymbol{\beta}_o$, instead of [A2](i). Under this condition,

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_T) = \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X})] = \boldsymbol{\beta}_o,$$

by Lemma 5.9 (law of iterated expectations). Note that the condition $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}_o$ implies

$$\mathbb{E}(\boldsymbol{y}) = \mathbb{E}[\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X})] = \mathbb{E}(\boldsymbol{X})\boldsymbol{\beta}_o,$$

again by the law of iterated expectations. Hence, $\mathbb{E}(\boldsymbol{y})$ can be obtained from $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X})$ but not conversely. This shows that, when $\boldsymbol{X}$ is allowed to be stochastic, the unbiasedness property of $\hat{\boldsymbol{\beta}}_T$ would hold under a stronger condition.

Unfortunately, the condition $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}_o$ may not be realistic in some applications. To see this, let $\boldsymbol{x}_t$ denote the t^{th} column of $\boldsymbol{X}'$ and write the t^{th} element of $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}_o$ as

$$\mathbb{E}(y_t \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T) = \boldsymbol{x}_t'\boldsymbol{\beta}_o, \qquad t = 1, 2, \ldots, T.$$

Consider time series data and the simple specification that $\boldsymbol{x}_t$ contains only one regressor y_{t-1}:

$$y_t = \beta y_{t-1} + e_t, \qquad t = 1, 2, \ldots, T.$$

In this case, the aforementioned condition reads:

$$\mathbb{E}(y_t \mid y_1, \ldots, y_{T-1}) = \beta_o y_{t-1},$$

for some β_o. Note that for $t = 1, \ldots, T-1$, $\mathbb{E}(y_t \mid y_1, \ldots, y_{T-1}) = y_t$ by Lemma 5.10. The condition above then requires $y_t = \beta_o y_{t-1}$ with probability one. If $\{y_t\}$ is indeed an AR(1) process: $y_t = \beta_o y_{t-1} + \epsilon_t$ such that ϵ_t has a continuous distribution, the event that $y_t = \beta_o y_{t-1}$ (i.e., $\epsilon_t = 0$) can occur only with probability zero, violating the imposed condition.

Suppose that $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}_o$ and $\operatorname{var}(\boldsymbol{y} \mid \boldsymbol{X}) = \sigma_o^2\boldsymbol{I}_T$. It is easy to see that

$$\begin{aligned}
\operatorname{var}(\hat{\boldsymbol{\beta}}_T) &= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_o)'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}] \\
&= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\operatorname{var}(\boldsymbol{y} \mid \boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}] \\
&= \sigma_o^2\,\mathbb{E}(\boldsymbol{X}'\boldsymbol{X})^{-1},
\end{aligned}$$

which is different from the variance-covariance matrix when $\boldsymbol{X}$ is non-stochastic; cf. Theorem 3.4(c). It is not always reasonable to impose such a condition on $\operatorname{var}(\boldsymbol{y} \mid \boldsymbol{X})$

either. Consider the previous example that $\boldsymbol{x}_t = y_{t-1}$. As $\mathbb{E}(y_t \mid y_1, \ldots, y_{T-1}) = y_t$, the conditional variance is

$$\operatorname{var}(y_t \mid y_1, \ldots, y_{T-1}) = \mathbb{E}\{[y_t - \mathbb{E}(y_t \mid y_1, \ldots, y_{T-1})]^2 \mid y_1, \ldots, y_{T-1}\} = 0,$$

rather than a positive constant σ_o^2.

Without the conditions on $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X})$ and $\operatorname{var}(\boldsymbol{y} \mid \boldsymbol{X})$, the mean and variance of the OLS estimator remain unknown. Moreover, when $\boldsymbol{X}$ is stochastic, $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ need not be normally distributed even when $\boldsymbol{y}$ is. Consequently, the results for hypothesis testing discussed in Section 3.4 are invalid.

6.2 Asymptotic Properties of the OLS Estimators

Suppose that we observe the data $(y_t \; \boldsymbol{w}_t')'$, where y_t is the variable of interest (dependent variable), and $\boldsymbol{w}_t$ is an $m \times 1$ vector of "exogenous" variables. Let $\mathcal{W}^t$ denote the collection of random vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_t$ and $\mathcal{Y}^t$ the collection of $y_1, \ldots, y_t$. The set of $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$ generates a σ-algebra which represents the information set up to time t. To account for the behavior of y_t, we choose the vector of explanatory variables $\boldsymbol{x}_t$ from the information set so that $\boldsymbol{x}_t$ includes k elements of $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$. The linear specification $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ can be expressed as

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t, \qquad t = 1, 2, \ldots, T, \tag{6.1}$$

where $\boldsymbol{x}_t$ is the t^{th} column of $\boldsymbol{X}'$, i.e., the t^{th} observation of all explanatory variables. Under the present framework, regressors may be lagged dependent variables (taken from $\mathcal{Y}^{t-1}$) and lagged exogenous variables (taken from $\mathcal{W}^t$). Including such variables in the specification is quite helpful in capturing the dynamic behavior of data.

6.2.1 Consistency

Gien the specification (6.1), the OLS estimator can be written as

$$\hat{\boldsymbol{\beta}}_T = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \left(\sum_{t=1}^{T} \boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{x}_t y_t\right). \tag{6.2}$$

The estimator $\hat{\boldsymbol{\beta}}_T$ is said to be *strongly consistent* for the parameter vector $\boldsymbol{\beta}^*$ if $\hat{\boldsymbol{\beta}}_T \xrightarrow{\text{a.s.}} \boldsymbol{\beta}^*$ as T tends to infinity; $\hat{\boldsymbol{\beta}}_T$ is said to be *weakly consistent* for $\boldsymbol{\beta}^*$ if $\hat{\boldsymbol{\beta}}_T \xrightarrow{\mathbb{P}} \boldsymbol{\beta}^*$. Strong consistency asserts that $\hat{\boldsymbol{\beta}}_T$ will be eventually close to $\boldsymbol{\beta}^*$ when "enough" information (a sufficiently large sample) becomes available. Consistency is in sharp contrast with

unbiasedness. While an unbiased estimator of $\boldsymbol{\beta}^*$ is "correct" on average, there is no guarantee that its values will be close to $\boldsymbol{\beta}^*$, no matter how large the sample is.

To establish strong (weak) consistency, we impose the following conditions.

[B1] $\{(y_t \ \boldsymbol{w}_t')'\}$ is a sequence of random vectors and $\boldsymbol{x}_t$ is also a random vector containing some elements of $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$.

(i) $\{\boldsymbol{x}_t\boldsymbol{x}_t'\}$ obeys a SLLN (WLLN) such that $\lim_{T\to\infty} T^{-1}\sum_{t=1}^{T} \mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')$ exists and is nonsingular.

(ii) $\{\boldsymbol{x}_t y_t\}$ obeys a SLLN (WLLN).

[B2] For some $\boldsymbol{\beta}_o$, $\mathbb{E}(y_t \mid \mathcal{Y}^{t-1}, \mathcal{W}^t) = \boldsymbol{x}_t'\boldsymbol{\beta}_o$ for all t.

One approach in the time-series analysis is to analyze the behavior of y_t based solely on its past behavior (lagged values). In this case, $\boldsymbol{x}_t$ contains only the elements of $\mathcal{Y}^{t-1}$, and [B2] is modifed as $\mathbb{E}(y_t \mid \mathcal{Y}^{t-1}) = \boldsymbol{x}_t'\boldsymbol{\beta}_o$ for all t.

The condition [B1] explicitly allows the explanatory variables $\boldsymbol{x}_t$ to be a random vector which may contain one or more lagged dependent variables y_{t-j} and current and past exogenous variables $\boldsymbol{w}_t$. [B1] also admits non-stochastic regressors which can be viewed as independent, degenerate random vectors. Moreover, [B1](i) and (ii) regulate the behaviors of y_t and $\boldsymbol{x}_t$ such that $\{\boldsymbol{x}_t\boldsymbol{x}_t'\}$ and $\{\boldsymbol{x}_t y_t\}$ must obey a SLLN (WLLN). On the other hand, the deterministic time trend t and random walk are excluded because they do not obey a SLLN (WLLN); see Examples 5.29 and 5.31.

Analogous to [A2](i), [B2] requires the linear function $\boldsymbol{x}_t'\boldsymbol{\beta}$ to be a correct specification of the conditional mean function, up to some unknown parameters. When $\boldsymbol{x}_t$ is non-stochastic, [B2] implies [A2](i) because by the law of iterated expectations,

$$\mathbb{E}(y_t) = \mathbb{E}[\mathbb{E}(y_t \mid \mathcal{Y}^{t-1}, \mathcal{W}^t)] = \boldsymbol{x}_t'\boldsymbol{\beta}_o.$$

Recall from Section 5.2 that the conditional mean $\mathbb{E}(y_t \mid \mathcal{Y}^{t-1}, \mathcal{W}^t)$ is the orthogonal projection of y_t onto the space of all functions of the elements of $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$, where orthogonality is defined in terms of the cross moment, the inner product in L_2 space. Thus, the conditional mean function is the best approximation (in the mean squared error sense) of y_t based on the information generated by $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$.

As $\boldsymbol{x}_t'\boldsymbol{\beta}_o$ is the orthogonal projection of y_t under [B2], it must be true that, for any function g and any vector $\boldsymbol{z}_t$ containing the elements of $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$,

$$\mathbb{E}[g(\boldsymbol{z}_t)(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_o)] = 0,$$

by Lemma 5.11. That is, any function of $\boldsymbol{z}_t$ must be orthogonal to the difference between y_t and its orthogonal projection $\boldsymbol{x}_t'\boldsymbol{\beta}_o$. If this condition does not hold for some $g(\boldsymbol{z}_t)$, it should be clear that $\boldsymbol{x}_t'\boldsymbol{\beta}_o$ cannot be $\mathbb{E}(y_t \mid \mathcal{Y}^{t-1}, \mathcal{W}^t)$. In particular, if

$$\mathbb{E}[\boldsymbol{x}_t(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_o)] \neq \mathbf{0},$$

$\boldsymbol{x}_t'\boldsymbol{\beta}_o$ cannot be the conditional mean.

Unlike [A2](ii), the imposed conditions do not rule out serially correlated y_t, nor do they require the conditional variance $\operatorname{var}(y_t \mid \mathcal{Y}^{t-1}, \mathcal{W}^t)$ to be a constant. Moreover, $\{\boldsymbol{x}_t\}$ may also be a sequence of weakly dependent and heterogeneously distributed random variables, as long as it obeys a SLLN (WLLN). To summarize, the conditions here allow data to exhibit various forms of dependence and heterogeneity. By contrast, the classical conditions admit only serially uncorrelated and homoskedastic data.

Given [B1], define the following limits:

$$\boldsymbol{M}_{xx} := \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t'), \qquad \boldsymbol{M}_{xy} := \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}(\boldsymbol{x}_t y_t),$$

which are, respectively, the almost surely (probability) limits of the average of $\boldsymbol{x}_t\boldsymbol{x}_t'$ and $\boldsymbol{x}_t y_t$ under a SLLN (WLLN). As matrix inversion is a continuous function and $\boldsymbol{M}_{xx}$ is invertible by [B1](i), Lemma 5.13 (Lemma 5.17) ensures that

$$\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1} \to \boldsymbol{M}_{xx}^{-1} \quad \text{a.s. (in probability)}.$$

It follows from (6.2) that

$$\hat{\boldsymbol{\beta}}_T = \left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t y_t\right) \to \boldsymbol{M}_{xx}^{-1}\boldsymbol{M}_{xy} \quad \text{a.s. (in probability)}.$$

This shows that the OLS estimator of $\boldsymbol{\beta}$ has a well-defined limit under [B1].

Theorem 6.1 *Given the linear specification* (6.1)*, suppose that* [B1] *holds. Then,* $\hat{\boldsymbol{\beta}}_T$ *is srongly (weakly) consistent for* $\boldsymbol{M}_{xx}^{-1}\boldsymbol{M}_{xy}$*.*

Theorem 6.1 holds regardless of [B2]; that is, whether (6.1) is the correct specification or not is irrelevant.

Example 6.2 Given the simple AR(1) specification

$$y_t = \alpha y_{t-1} + e_t,$$

suppose that $\{y_t^2\}$ and $\{y_t y_{t-1}\}$ obey a SLLN (WLLN). Then, Theorem 6.1 ensures that the OLS estimator

$$\hat{\alpha}_T \to \frac{\lim_{T\to\infty} \frac{1}{T}\sum_{t=2}^{T} \mathbb{E}(y_t y_{t-1})}{\lim_{T\to\infty} \frac{1}{T}\sum_{t=2}^{T} \mathbb{E}(y_{t-1}^2)} \quad \text{a.s. (in probability).}$$

If $\mathbb{E}(y_t) = 0$, $\mathrm{var}(y_t) = \gamma_0$ and $\mathrm{cov}(y_t, y_{t-1}) = \gamma_1$ for all t. This limit simplifies to γ_1/γ_0.

By the law of iterated expectations, [B2] implies

$$\mathbb{E}(\boldsymbol{x}_t y_t) = \mathbb{E}[\boldsymbol{x}_t\, \mathbb{E}(y_t \mid \mathcal{Y}^{t-1}, \mathcal{W}^t)] = \mathbb{E}(\boldsymbol{x}_t \boldsymbol{x}_t')\boldsymbol{\beta}_o,$$

which in turn yields $\boldsymbol{M}_{xy} = \boldsymbol{M}_{xx}\boldsymbol{\beta}_o$. Thus, [B1] and [B2] together determine the almost sure (probability) limit in Theorem 6.1 as

$$\boldsymbol{M}_{xx}^{-1}\boldsymbol{M}_{xy} = \boldsymbol{\beta}_o.$$

We have established the following result.

Theorem 6.3 *Given the linear specification* (6.1)*, suppose that* [B1] *and* [B2] *hold. Then, $\hat{\boldsymbol{\beta}}_T$ is strongly (weakly) consistent for the parameter vector $\boldsymbol{\beta}_o$ in* [B2]*.*

We state consistency in two results so as to clarify the convergence behavior of the OLS estimator. First, as long as the data obey a proper SLLN (WLLN), $\hat{\boldsymbol{\beta}}_T$ converges to "something." Second, the almost sure (probability) limit would be $\boldsymbol{\beta}_o$ if $\boldsymbol{x}_t'\boldsymbol{\beta}$ is the correct specification for the conditional mean. These results are convenient for asserting consistency. Once we believe (or are able to verify) that the data obey a SLLN (WLLN), the conclusion of Theorem 6.1 immediately applies. If, further, we believe (or are able to verify) that the specification is correct for the conditional mean, we can conclude from Theorem 6.3 that the OLS estimator is strongly (weakly) consistent for the parameter of interest.

Example 6.4 Given the simple AR(1) specification

$$y_t = \alpha y_{t-1} + e_t,$$

suppose that

$$y_t = \alpha_o y_{t-1} + u_t,$$

where $|\alpha_o| < 1$ and $\{u_t\}$ is a sequence of unobservable, independent random variables with mean zero and variance σ_u^2. A process so generated is an AR(1) process. As all the elements of $\mathcal{Y}^{t-1}$ are determined by u_s for $s \leq t-1$, it is then clear that these elements and their functions must be independent of u_t, by Lemma 5.1. It follows that $\alpha_o y_{t-1}$ is the conditional mean $\mathbb{E}(y_t \mid \mathcal{Y}^{t-1})$. Theorem 6.3 now ensures that $\hat{\alpha}_T \to \alpha_o$ a.s. (in probability). Note, however, that $\alpha_o y_{t-1}$ need not be the conditional mean if $\{u_t\}$ is a white noise sequence.

Alternatively, we can establish consistency as follows. In view of Section 4.4, y_t is weakly stationary with mean zero, variance $\sigma_u^2/(1-\alpha_o^2)$ and

$$\operatorname{cov}(y_t, y_{t-j}) = \alpha_o^j \frac{\sigma_u^2}{1-\alpha_o^2}.$$

It follows from Example 6.2 that

$$\hat{\alpha}_T \to \frac{\operatorname{cov}(y_t, y_{t-1})}{\operatorname{var}(y_{t-1})} = \alpha_o \quad \text{a.s. (in probability)}.$$

Comparing to Example 6.2 we can see that the more we know about data, the more precise we can say about the limit of the OLS estimator. □

The examples below illustrate that when $\boldsymbol{x}_t'\boldsymbol{\beta}_o$ is not the desired conditional mean, the OLS estimator still converges but may be inconsistent for $\boldsymbol{\beta}_o$.

Example 6.5 Consider the specification

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t,$$

where $\boldsymbol{x}_t'$ is $k_1 \times 1$. Suppose that

$$\mathbb{E}(y_t \mid \mathcal{Y}^{t-1}, \mathcal{W}^t) = \boldsymbol{x}_t'\boldsymbol{\beta}_o + \boldsymbol{z}_t'\boldsymbol{\gamma}_o,$$

where $\boldsymbol{z}_t$ ($k_2 \times 1$) also contains the elements of $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$ and is distinct from $\boldsymbol{x}_t$. This is an example that a specification omits relevant variables ($\boldsymbol{z}_t$ in the conditional mean). When [B1] holds,

$$\hat{\boldsymbol{\beta}}_T \to \boldsymbol{M}_{xx}^{-1}\boldsymbol{M}_{xy}, \quad \text{a.s. (in probability)},$$

by Theorem 6.1. In this case,

$$\mathbb{E}(\boldsymbol{x}_t y_t) = \mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')\boldsymbol{\beta}_o + \mathbb{E}(\boldsymbol{x}_t\boldsymbol{z}_t')\boldsymbol{\gamma}_o,$$

so that

$$\mathbb{E}[\boldsymbol{x}_t(y_t - \boldsymbol{x}_t\boldsymbol{\beta}_o)] = \mathbb{E}(\boldsymbol{x}_t\boldsymbol{z}_t')\boldsymbol{\gamma}_o,$$

which is non-zero in general. This shows that $\boldsymbol{x}_t'\boldsymbol{\beta}_o$ cannot be the desired conditional mean, and hence convergence to $\boldsymbol{\beta}_o$ is not guaranteed. In fact, the almost sure (probability) limit of $\hat{\boldsymbol{\beta}}_T$ is

$$\boldsymbol{\beta}^* = \boldsymbol{M}_{xx}^{-1}\boldsymbol{M}_{xy} = \boldsymbol{\beta}_o + \boldsymbol{M}_{xx}^{-1}\boldsymbol{M}_{xz}\boldsymbol{\gamma}_o,$$

where $\boldsymbol{M}_{xz} = \lim_{T\to\infty}\sum_{t=1}^T \mathbb{E}(\boldsymbol{x}_t\boldsymbol{z}_t')/T$ is assumed to exist. This limit is not $\boldsymbol{\beta}_o$ in general. If the elements of $\boldsymbol{x}_t$ are orthogonal to those of $\boldsymbol{z}_t$, i.e., $\mathbb{E}(\boldsymbol{x}_t\boldsymbol{z}_t') = \boldsymbol{0}$, we have $\boldsymbol{M}_{xz} = \boldsymbol{0}$ and $\boldsymbol{\beta}^* = \boldsymbol{\beta}_o$. Then $\hat{\boldsymbol{\beta}}_T$ converges to $\boldsymbol{\beta}_o$ almost surely (in probability), despite that the linear function is misspecified for the conditional mean function. □

Example 6.6 Given the simple AR(1) specification

$$y_t = \alpha y_{t-1} + e_t,$$

suppose that

$$y_t = \alpha_o y_{t-1} + u_t + \pi_o u_{t-1},$$

where $|\alpha_o| < 1$, $|\pi_o| < 1$, and $\{u_t\}$ is a sequence of unobservable, independent random variables with mean zero and variance σ_u^2. A process so generated is known as an ARMA(1,1) process (autoregressive and moving average process of order (1,1)). It is easily shown that $\alpha_o y_{t-1}$ is not the conditional mean $\mathbb{E}(y_t \mid \mathcal{Y}^{t-1})$ because

$$\mathbb{E}[y_{t-1}(y_t - \alpha_o y_{t-1})] = \mathbb{E}[y_{t-1}(u_t + \pi_o u_{t-1})] = \pi_o\sigma_u^2,$$

which is non-zero unless $\pi_o = 0$. As $\{y_t\}$ is weakly stationary such that

$$\mathbb{E}(y_t y_{t-1}) = \alpha_o\,\mathbb{E}(y_{t-1}^2) + \pi_o\sigma_u^2,$$

where $\mathbb{E}(y_{t-1}^2)$ is a constant, we have from Example 6.2 that

$$\hat{\alpha}_T \to \frac{\mathbb{E}(y_t y_{t-1})}{\mathbb{E}(y_{t-1}^2)} = \alpha_o + \pi_o\frac{\sigma_u^2}{\mathbb{E}(y_{t-1}^2)} \quad \text{a.s. (in probability)}.$$

This shows that $\hat{\alpha}_T$ converges but is inconsistent for α_o. This is the case because $y_t - \alpha_o y_{t-1}$ are serially correlated and hence correlated with y_{t-1}. Clearly, $\hat{\alpha}_T$ would be consistent for α_o if $\pi_o = 0$, i.e., $y_t - \alpha_o y_{t-1}$ are serially uncorrelated (in fact, independent) and hence uncorrelated with y_{t-1}, as shown in Example 6.4.

This conclusion will not be altered if the lagged dependent variable is one of the regressors:

$$y_t = \alpha y_{t-1} + \boldsymbol{x}_t'\boldsymbol{\beta} + e_t.$$

Suppose that

$$y_t = \alpha_o y_{t-1} + \boldsymbol{x}_t'\boldsymbol{\beta}_o + \epsilon_t,$$

where ϵ_t are serially correlated. Then the OLS estimators for α and $\boldsymbol{\beta}$ are inconsistent for α_o and $\boldsymbol{\beta}_o$. □

In the examples above, y_t can be written as

$$y_t = \boldsymbol{x}_t\boldsymbol{\beta}^* + \epsilon_t,$$

where ϵ_t are disturbances. For Example 6.5, $\boldsymbol{\beta}^* = \boldsymbol{\beta}_o$, $\epsilon_t = \boldsymbol{z}_t'\boldsymbol{\gamma}_o + u_t$, and u_t is such that $\mathbb{E}(u_t \mid \mathcal{Y}^{t-1}, \mathcal{W}^t) = 0$; for Example 6.6, $\boldsymbol{\beta}^* = \alpha_o$, $\epsilon_t = u_t + \gamma_o u_{t-1}$, and $\{u_t\}$ is a white noise sequence. By noting

$$\hat{\boldsymbol{\beta}}_T = \boldsymbol{\beta}^* + \left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}^*)\right), \tag{6.3}$$

we can see that $\hat{\boldsymbol{\beta}}_T$ would converge to $\boldsymbol{\beta}^*$ provided that the second term on the right-hand side vanishes in the limit. When $\boldsymbol{x}_t$ are uncorrelated with the disturbances: $\mathbb{E}[\boldsymbol{x}_t(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}^*)] = \boldsymbol{0}$ for all t, SLLN (WLLN) implies

$$\frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}^*) \to \boldsymbol{0} \quad \text{a.s. (in probability)},$$

$$\frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t' = O(1) \quad \text{a.s. (in probability)}.$$

It follows that the second term on the right-hand side of (6.3) does converge to zero almost surely (in probability). Clearly, [B2] is a sufficient condition for $\mathbb{E}[\boldsymbol{x}_t(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}^*)] = \boldsymbol{0}$ for $\boldsymbol{\beta}^* = \boldsymbol{\beta}_o$.

On the other hand, when $\mathbb{E}[\boldsymbol{x}_t(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}^*)] = \boldsymbol{c} \neq \boldsymbol{0}$,

$$\frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}^*) \to \boldsymbol{c} \quad \text{a.s. (in probability)},$$

so that $\hat{\boldsymbol{\beta}}_T$ converges in probability to $\boldsymbol{\beta}^* + \boldsymbol{M}_{xx}^{-1}\boldsymbol{c}$, which differs from $\boldsymbol{\beta}^*$ by a fixed amount. That is, inconsistency would result if the regressor $\boldsymbol{x}_t$ are correlated with the disturbances ϵ_t. Such correlations may be due to the correlation between the included and excluded variables (Example 6.5) or the correlation between the lagged dependent variables and serially correlated disturbances (Example 6.6).

While the effect of SLLN or WLLN (condition [B1]) is important in establishing OLS consistency, it will be shown below that [B1] is not a necessary condition.

Example 6.7 Given the simple linear time trend specification:

$$y_t = a + b\,t + e_t,$$

suppose that [B2] holds: $\mathbb{E}(y_t|\mathcal{Y}^{t-1}) = a_o + b_o\,t$. We have learned from Example 5.29 that $\{t\}$ and $\{t^2\}$ do not obey a SLLN or a WLLN so that [B1] is violated. Nevertheless, the OLS estimators of a and b remain consistent. In view of (6.3), the OLS estimator of b is

$$\hat{b}_T = \frac{\sum_{t=1}^T \left(t - \frac{T+1}{2}\right) y_t}{\sum_{t=1}^T \left(t - \frac{T+1}{2}\right)^2} = b_o + \frac{\sum_{t=1}^T \left(t - \frac{T+1}{2}\right) \epsilon_t}{\sum_{t=1}^T \left(t - \frac{T+1}{2}\right)^2},$$

where $\epsilon_t = y_t - a_o - b_o t$. We have seen in Example 5.30 that $\sum_{t=1}^T \epsilon_t$ is $O_{\mathbb{P}}(T^{1/2})$ and $\sum_{t=1}^T t\epsilon_t$ is $O_{\mathbb{P}}(T^{3/2})$. While the numerator term is $O_{\mathbb{P}}(T^{3/2})$, the denominator grows even faster:

$$\sum_{t=1}^T \left(t - \frac{T+1}{2}\right)^2 = \sum_{t=1}^T t^2 - \frac{T(T+1)^2}{4} = \frac{T(T+1)(T-1)}{12} = O(T^3).$$

The entire second term thus vanishes in the limit, and $\hat{b}_T \xrightarrow{\mathbb{P}} b_o$. Similarly, we can show

$$\hat{a}_T = \bar{y} - \hat{b}_T \frac{T+1}{2} = a_o + \frac{(4T+2)\sum_{t=1}^T \epsilon_t - 6\sum_{t=1}^T t\epsilon_t}{T(T-1)}.$$

As the second term above is $O_{\mathbb{P}}(T^{-1/2})$, $\hat{a}_T \xrightarrow{\mathbb{P}} a_o$. □

Example 6.8 Given a simple AR(1) specification:

$$y_t = \alpha y_{t-1} + e_t,$$

suppose that y_t is a random walk:

$$y_t = y_{t-1} + \epsilon_t,$$

where ϵ_t are i.i.d. random variables. We have seen in Example 5.31 that $\{y_t\}$ and $\{y_t^2\}$ do not obey a SLLN (WLLN). By (6.3), the OLS estimator of α can be written as

$$\hat{\alpha}_T = 1 + \frac{\sum_{t=1}^T y_{t-1}\epsilon_t}{\sum_{t=1}^T y_{t-1}^2}.$$

From Examples 5.31 and 5.32 we know that the numerator on the right-hand side above is $O_{\mathbb{P}}(T)$, while the denominator is $O_{\mathbb{P}}(T^2)$. Consequently, $\hat{\alpha}_T \xrightarrow{\mathbb{P}} 1$.

When $\{\epsilon_t\}$ is a weakly stationary ARMA process and exhibits serial correlations, y_{t-1} is not the conditional mean of y_t because $\mathbb{E}(y_{t-1}\epsilon_t)$ is non-zero. Nevertheless,

$$\frac{\sum_{t=1}^T y_{t-1}\epsilon_t}{\sum_{t=1}^T y_{t-1}^2} = \frac{O_{\mathbb{P}}(T)}{O_{\mathbb{P}}(T^2)} = O_{\mathbb{P}}(T^{-1}), \tag{6.4}$$

so that $\hat{\alpha}_T$ is still weakly consistent for 1. □

Remark: Example 6.8 demonstrates that the OLS estimator may still be consistent even when a lagged dependent variable and serially correlated disturbances are both present. This is because $\sum_{t=1}^T y_{t-1}^2$ in (6.4) grows much faster and hence is able to eliminate all the correlations between y_{t-1} and ϵ_t asymptotically. If $\sum_{t=1}^T y_{t-1}^2$ and $\sum_{t=1}^T y_{t-1}\epsilon_t$ in (6.4) grow at the same rate, these correlations would not vanish in the limit and therefore cause inconsistency, as shown in Example 6.6.

6.2.2 Asymptotic Normality

We say that $\hat{\boldsymbol{\beta}}_T$ is *asymptotically normally distributed* (about $\boldsymbol{\beta}^*$) if

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{D}_o),$$

where $\boldsymbol{D}_o$ is a positive-definite matrix. That is, the sequence of properly normalized $\hat{\boldsymbol{\beta}}_T$ converges in distribution to a multivariate normal random vector. As $\boldsymbol{D}_o$ is the covariance matrix of the limiting normal distribution, it is also known as the *asymptotic covariance matrix* of $\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*)$. Equivalently, we may also express asymptotic normality by

$$\boldsymbol{D}_o^{-1/2}\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{I}_k).$$

It should be emphasized that asymptotic normality here is referred to $\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*)$ rather than $\hat{\boldsymbol{\beta}}_T$; the latter has only a degenerate distribution in the limit by strong (weak) consistency.

When $\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*)$ has a limiting distribution, it is $O_{\mathbb{P}}(1)$ by Lemma 5.24. Therefore, $\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*$ is necessarily $O_{\mathbb{P}}(T^{-1/2})$, so that $\hat{\boldsymbol{\beta}}_T$ tend to $\boldsymbol{\beta}^*$ at the rate $T^{-1/2}$. Thus, we know not only consistency but also the rate of convergence to $\boldsymbol{\beta}^*$. An estimator that is consistent at the rate $T^{-1/2}$ is usually referred to as a "$\sqrt{T}$-consistent" estimator. Some consistent estimators may converge more quickly. In Example 6.7, the estimator $\hat{b}_T$ of the slope coefficient in the simple time trend specification converges to b_o at the rate $T^{-3/2}$, whereas the estimator of the intercept is $\sqrt{T}$-consistent. Also, the OLS estimator for the AR(1) specification is T-consistent when $\{y_t\}$ is a random walk but $\sqrt{T}$-consistent when $\{y_t\}$ is a weakly stationary process; see Examples 6.8.

To ensure asymptotic normality, we impose an additional condition.

[B3] For some $\boldsymbol{\beta}^*$, $\{\boldsymbol{x}_t(y_t - \boldsymbol{x}_t'\boldsymbol{\beta}^*)\}$ is a sequence of random vectors with mean zero and obeys a CLT.

If we write

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta}^* + \epsilon_t,$$

[B3] requires that $\mathbb{E}(\boldsymbol{x}_t\epsilon_t) = \boldsymbol{0}$, i.e., the regressors $\boldsymbol{x}_t$ and disturbances ϵ_t are uncorrelated. Moreover,

$$\boldsymbol{V}_T := \operatorname{var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{x}_t\epsilon_t\right) \to \boldsymbol{V}_o,$$

a positive-definite matrix, and

$$\boldsymbol{V}_o^{-1/2}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{x}_t\epsilon_t\right) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_k).$$

In view of (6.3), the normalized OLS estimator is

$$\begin{aligned}\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) &= \left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{x}_t\epsilon_t\right) \\ &= \left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\boldsymbol{V}_o^{1/2}\left[\boldsymbol{V}_o^{-1/2}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{x}_t\epsilon_t\right)\right].\end{aligned} \tag{6.5}$$

Given the SLLN (WLLN) condition [B1] and the CLT condition [B3], the first term on the right-hand side of (6.5) converges in probability to $\boldsymbol{M}_{xx}^{-1}$, and the last term in the

square bracket converges in distribution to $N(\mathbf{0}, \boldsymbol{I}_k)$. It follows from Lemma 5.22 that (6.5) converges in distribution to

$$\boldsymbol{M}_{xx}^{-1}\boldsymbol{V}_o^{1/2}\, N(\mathbf{0}, \boldsymbol{I}_k) \overset{d}{=} N(\mathbf{0},\, \boldsymbol{M}_{xx}^{-1}\boldsymbol{V}_o\boldsymbol{M}_{xx}^{-1}),$$

where $\overset{d}{=}$ stands for equality in distribution. We have established the following asymptotic normality result.

Theorem 6.9 *Given the linear specification* (6.1)*, suppose that* [B1] *and* [B3] *hold. Then,*

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{D}_o),$$

where $\boldsymbol{D}_o = \boldsymbol{M}_{xx}^{-1}\boldsymbol{V}_o\boldsymbol{M}_{xx}^{-1}$*, or equivalently,*

$$\boldsymbol{D}_o^{-1/2}\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{I}_k),$$

where $\boldsymbol{D}_o^{-1/2} = \boldsymbol{V}_o^{-1/2}\boldsymbol{M}_{xx}$.

As long as we believe (or are able to verify) that the data have proper SLLN (WLLN) and CLT effects, we can invoke Theorem 6.9 and assert asymptotic normality of the (properly normalized) OLS estimator. In particular, this result may hold for weakly dependent and heterogeneously distributed data; neither independence nor normality is required. By contrast, the normality property in Theorem 3.7(a) is an exact distribution result for the OLS estimator, but it is valid only when y_t are are independent, normal random variables.

When $\boldsymbol{V}_o$ is unknown, let $\hat{\boldsymbol{V}}_T$ denote a symmetric and positive definite matrix that is consistent for $\boldsymbol{V}_o$. A weakly consistent estimator of $\boldsymbol{D}_o$ is then

$$\hat{\boldsymbol{D}}_T = \left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\hat{\boldsymbol{V}}_T\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}.$$

As $\hat{\boldsymbol{D}}_T \xrightarrow{\mathbb{P}} \boldsymbol{D}_o$, we must have $\hat{\boldsymbol{D}}_T^{-1/2} \xrightarrow{\mathbb{P}} \boldsymbol{D}_o^{-1/2}$. It follows from Theorem 6.9 and Lemma 5.19 that

$$\hat{\boldsymbol{D}}_T^{-1/2}\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) \xrightarrow{D} \boldsymbol{D}_o^{-1/2}\, N(\mathbf{0}, \boldsymbol{D}_o) \overset{d}{=} N(\mathbf{0}, \boldsymbol{I}_k).$$

The shows that Theorem 6.9 remains valid when the asymptotic covariance matrix $\boldsymbol{D}_o$ is replaced by a weakly consistent estimator $\hat{\boldsymbol{D}}_T$. Note that $\hat{\boldsymbol{D}}_T$ does not have to be a strongly consistent estimator here.

Theorem 6.10 *Given the linear specification* (6.1)*, suppose that* [B1] *and* [B3] *hold. Then,*

$$\hat{\boldsymbol{D}}_T^{-1/2}\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_k),$$

where $\hat{\boldsymbol{D}}_T = (\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t'/T)^{-1}\hat{\boldsymbol{V}}_T(\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t'/T)^{-1}$ *and* $\hat{\boldsymbol{V}}_T \xrightarrow{\mathbb{P}} \boldsymbol{V}_o$.

Remark: Theorems 6.9 and 6.10 show that we may obtain asymptotic normality when the condition [B2] for correct specification is absent. Without [B2], there is no guarantee that $\hat{\boldsymbol{\beta}}_T$ would converge to $\boldsymbol{\beta}_o$, but it still converges to some limit $\boldsymbol{\beta}^*$ under [B1]. Then the CLT effect of [B3] suffices for asymptotic normality. When the asymptotic covariance matrix $\boldsymbol{V}_o$ is unknown, it is of paramount importance to find a consistent estimator of $\boldsymbol{D}_o$. Normalizing the OLS estimator with an inconsistent estimator of $\boldsymbol{D}_o$ will, in general, destroy asymptotic normality.

Example 6.11 Given the linear specification

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e, \qquad t = 1, \ldots, T,$$

suppose that the classical conditions [A1] and [A2] hold. If $\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t'/T$ converges to some limit $\boldsymbol{M}_{xx}$, then

$$\boldsymbol{V}_o = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \mathbb{E}(\epsilon_t^2\boldsymbol{x}_t\boldsymbol{x}_t') = \sigma_o^2\boldsymbol{M}_{xx}.$$

By invoking a suitable CLT, it can be shown that the classical conditions are sufficient for [B3]. It follows from Theorem 6.9 that

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{D}_o),$$

with a much simpler asymptotic covariance matrix:

$$\boldsymbol{D}_o = \boldsymbol{M}_{xx}^{-1}\boldsymbol{V}_o\boldsymbol{M}_{xx}^{-1} = \sigma_o^2\boldsymbol{M}_{xx}^{-1}.$$

Comparing to Theorem 3.7(a), y_t here need not be normally distributed, and the asymptotic covariance matrix is determined by the limit of $\sum_{t=1}^T(\boldsymbol{x}_t\boldsymbol{x}_t')/T$, which in matrix notations can be written as $\boldsymbol{X}'\boldsymbol{X}/T$. A natural estimator of $\boldsymbol{D}_o$ is

$$\hat{\boldsymbol{D}}_T = \hat{\sigma}_T^2\,(\boldsymbol{X}'\boldsymbol{X}/T)^{-1},$$

where $\hat{\sigma}_T^2$ is the OLS variance estimator. Theorem 6.10 then ensures

$$\frac{1}{\hat{\sigma}_T}(\boldsymbol{X}'\boldsymbol{X}/T)^{1/2}\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) = \frac{1}{\hat{\sigma}_T}(\boldsymbol{X}'\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_k),$$

Taking the inner product of the left-hand side above we immediately conclude that

$$\frac{(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*)'(\boldsymbol{X}'\boldsymbol{X})(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o)}{\hat{\sigma}_T^2} \xrightarrow{D} \chi^2(k).$$

by Lemma 5.20. Note that the left-hand side is k times the F statistic (with $\boldsymbol{R} = \boldsymbol{I}_k$) in Section 3.4.1. □

The example below shows that even without the effects of SLLN (WLLN) and CLT, properly normalized OLS estimators may still have an asymptotic normal distribution.

Example 6.12 The simple linear time trend specification,

$$y_t = a + b\,t + e_t,$$

is a special case of the regression with non-stochastic regressors $\boldsymbol{x}_t = [1\ t]'$. Let $\hat{a}_T$ and $\hat{b}_T$ denote the OLS estimators of a and b, respectively. We know that $\{\boldsymbol{x}_t\boldsymbol{x}_t'\}$ does not obey a SLLN (WLLN) and that $\{t\epsilon_t\}$ does not obey a CLT. It is, however, easy to see that for $\tilde{\boldsymbol{x}}_t = [1\ t/T]'$,

$$\frac{1}{T}\sum_{t=1}^{T} \tilde{\boldsymbol{x}}_t\tilde{\boldsymbol{x}}_t' \to \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{bmatrix} =: \boldsymbol{M},$$

so that $\{\tilde{\boldsymbol{x}}_t\tilde{\boldsymbol{x}}_t'\}$ obeys a SLLN. Example 5.37 also shows that $\{(t/T)\epsilon_t\}$ obeys a CLT. These suggest that we may consider an alternative specification:

$$y_t = a + b\,\frac{t}{T} + e_t.$$

The resulting OLS estimators are such that $\tilde{a}_T = \hat{a}_T$ and $\tilde{b}_T = T\hat{b}_T$.

Suppose that

$$y_t = a_o + b_o\,t + \epsilon_t,$$

where ϵ_t are uncorrelated random variables with $\mathbb{E}(\epsilon_t) = 0$ and $\mathrm{var}(\epsilon_t) = \sigma_o^2$. In view of the preceding example we can then conclude that

$$\begin{bmatrix} T^{1/2}(\hat{a}_T - a_o) \\ T^{3/2}(\hat{b}_T - b_o) \end{bmatrix} = \sqrt{T}\begin{bmatrix} \tilde{a}_T - a_o \\ \tilde{b}_T - T\,b_o \end{bmatrix} \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{D}_o),$$

with $\boldsymbol{D}_o = \sigma_o^2\boldsymbol{M}^{-1}$, where

$$\boldsymbol{M}^{-1} = \begin{bmatrix} 4 & -6 \\ -6 & 12 \end{bmatrix}.$$

Moreover,

$$\boldsymbol{D}_o^{-1/2}\begin{bmatrix} T^{1/2}(\hat{a}_T - a_o) \\ T^{3/2}(\hat{b}_T - b_o) \end{bmatrix} \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_2),$$

where $\boldsymbol{D}_o^{-1/2} = \sigma_o^{-1}\boldsymbol{M}^{1/2}$ with

$$\boldsymbol{M}^{1/2} = \begin{bmatrix} 1 & 1/2 \\ 0 & 1/(2\sqrt{3}) \end{bmatrix}.$$

Replacing σ_o by the OLS estimator $\hat{\sigma}_T$ yields the same limiting result. □

6.3 Consistent Estimation of Covariance Matrix

We have seen in the preceding section that a consistent estimator of $\boldsymbol{D}_o = \boldsymbol{M}_{xx}^{-1}\boldsymbol{V}_o\boldsymbol{M}_{xx}^{-1}$ is crucial for the asymptotic normality result. The matrix $\boldsymbol{M}_{xx}$ can be consistently estimated by its sample counterpart $\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t'/T$; it then remains to find a consistent estimator of

$$\boldsymbol{V}_o = \lim_{T\to\infty} \operatorname{var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T \boldsymbol{x}_t\epsilon_t\right),$$

where $\epsilon_t = y_t - \boldsymbol{x}_t'\boldsymbol{\beta}^*$. This section is concerned with consistent estimation of $\boldsymbol{V}_o$ and $\boldsymbol{D}_o$.

In its most general form, $\boldsymbol{V}_o$ can be expressed as the sum of variances and autocovariances:

$$\begin{aligned}\boldsymbol{V}_o &= \lim_{T\to\infty} \operatorname{var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T \boldsymbol{x}_t\epsilon_t\right) \\ &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \mathbb{E}(\epsilon_t^2\boldsymbol{x}_t\boldsymbol{x}_t') + \\ &\qquad \lim_{T\to\infty}\frac{1}{T}\sum_{\tau=1}^{T-1}\sum_{t=\tau+1}^{T}\Big[\mathbb{E}(\boldsymbol{x}_{t-\tau}\epsilon_{t-\tau}\epsilon_t\boldsymbol{x}_t') + \mathbb{E}(\boldsymbol{x}_t\epsilon_t\epsilon_{t-\tau}\boldsymbol{x}_{t-\tau}')\Big].\end{aligned} \tag{6.6}$$

Let x_{ij} denote the j^{th} element of $\boldsymbol{x}_i$. It can be seen that for $t \neq s$,

$$\mathbb{E}(x_{t1}\epsilon_t\epsilon_s x_{s2}) \neq \mathbb{E}(x_{s1}\epsilon_s\epsilon_t x_{t2}),$$

in general. That is, the covariance matrix $\mathbb{E}(\boldsymbol{x}_{t-\tau}\epsilon_{t-\tau}\epsilon_t\boldsymbol{x}_t')$ need not be symmetric. This matrix would be symmetric when, for example, $\{\boldsymbol{x}_t\epsilon_t\}$ is a multivariate, weakly

stationary process such that the autocovariances of its elements, $\mathbb{E}(x_{ti}\epsilon_t\epsilon_s x_{sj})$, do not depend on t but only on the time difference $t-s$. When $\{\boldsymbol{x}_t\epsilon_t\}$ is indeed weakly stationary, $\boldsymbol{V}_o$ simplifies to

$$\boldsymbol{V}_o = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(\epsilon_t^2\boldsymbol{x}_t\boldsymbol{x}_t') + \lim_{T\to\infty}\frac{2}{T}\sum_{\tau=1}^{T-1}(T-\tau)\,\mathbb{E}(\boldsymbol{x}_{t-\tau}\epsilon_{t-\tau}\epsilon_t\boldsymbol{x}_t'). \tag{6.7}$$

Clearly, if $\boldsymbol{x}_t\epsilon_t$ are serially uncorrelated, the second terms on the right-hand side of (6.6) and (6.7) vanish; the remaining part of $\boldsymbol{V}_o$is relatively easy to estimate. When there are serial correlations, estimating $\boldsymbol{V}_o$ would be more cumbersome because it involves an infinite sum of autocovariances.

6.3.1 When Serial Correlations Are Absent

First observe that [B2] is equivalent to the condition that

$$\mathbb{E}(\epsilon_t|\mathcal{Y}^{t-1},\mathcal{W}^t) = 0,$$

where $\epsilon_t = y_t - \boldsymbol{x}_t\boldsymbol{\beta}_o$. The sequence $\{\epsilon_t\}$ with the property above is known as the *martingale difference sequence* with respect to the sequence of σ-algebras generated by $(\mathcal{Y}^{t-1},\mathcal{W}^t)$.

It is easy to see that if $\{\epsilon_t\}$ is a martingale difference sequence with respect to $\{\mathcal{Y}^{t-1},\mathcal{W}^t\}$, its unconditional mean and autocovariances are also zero, yet it may not be a white noise; see Exercise 6.7. Note also that a white noise need not be a martingale difference sequence. For the same reasons, we can verify that

$$\mathbb{E}(\boldsymbol{x}_t\epsilon_t) = \mathbb{E}[\boldsymbol{x}_t\,\mathbb{E}(\epsilon_t|\mathcal{Y}^{t-1},\mathcal{W}^t)] = \boldsymbol{0}.$$

and for any $t\neq\tau$,

$$\mathbb{E}(\boldsymbol{x}_t\epsilon_t\epsilon_\tau\boldsymbol{x}_\tau') = \mathbb{E}[\boldsymbol{x}_t\,\mathbb{E}(\epsilon_t|\mathcal{Y}^{t-1},\mathcal{W}^t)\epsilon_\tau\boldsymbol{x}_\tau'] = \boldsymbol{0}.$$

That is, $\{\boldsymbol{x}_t\epsilon_t\}$ is a sequence of uncorrelated, zero-mean random vectors under [B2]. In this case, the covariance matrices (6.6) and (6.7) are

$$\boldsymbol{V}_o = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(\epsilon_t^2\boldsymbol{x}_t\boldsymbol{x}_t'). \tag{6.8}$$

Note that the simpler form of $\boldsymbol{V}_o$ is a consequence of [B2], the correct specification of the conditional mean function. [B2] is not a necessary condition for (6.8), however.

If, in addition to [B2], ϵ_t are also *conditionally homoskedastic*:

$$\mathbb{E}(\epsilon_t^2|\mathcal{Y}^{t-1},\mathcal{W}^t)=\sigma_o^2,$$

then (6.8) can be further simplified to

$$\begin{aligned}\boldsymbol{V}_o &= \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathbb{E}(\epsilon_t^2|\mathcal{Y}^{t-1},\mathcal{W}^t)\boldsymbol{x}_t\boldsymbol{x}_t'] \\ &= \sigma_o^2\left(\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')\right) \\ &= \sigma_o^2\boldsymbol{M}_{xx}.\end{aligned} \tag{6.9}$$

As $\boldsymbol{M}_{xx}$ can be consistently estimated by its sample counterpart, it remains to estimate σ_o^2. It turns out that the standard OLS estimator $\hat{\sigma}_T^2=\sum_{t=1}^{T}\hat{e}_t^2/(T-k)$ is consistent for σ_o^2, where $\hat{e}_t$ are the OLS residuals; see Exercise 6.8. It follows that a consistent estimator of $\boldsymbol{V}_o$ is

$$\hat{\boldsymbol{V}}_T=\hat{\sigma}_T^2\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right).$$

In this case, the asymptotic covariance matrix of $\sqrt{T}(\hat{\beta}_T-\boldsymbol{\beta}_o)$ is also of a simpler form:

$$\boldsymbol{D}_o=\boldsymbol{M}_{xx}^{-1}\boldsymbol{V}_o\boldsymbol{M}_{xx}^{-1}=\sigma_o^2\boldsymbol{M}_{xx}^{-1},$$

which can be consistently estimated by

$$\hat{\boldsymbol{D}}_T=\hat{\sigma}_T^2\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}. \tag{6.10}$$

This is the same as the estimator obtained in Example 6.11. Note again that, apart from the factor T, $\hat{\boldsymbol{D}}_T$ here is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}_T$ in the classical least squares theory.

More generally, when ϵ_t are *conditionally heteroskedastic*, i.e., $\mathbb{E}(\epsilon_t^2|\mathcal{Y}^{t-1},\mathcal{W}^t)$ are random variables depending on t, then (6.8) cannot be simplified as before. To estimate (6.8), it can be seen that

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}[\hat{e}_t^2\boldsymbol{x}_t\boldsymbol{x}_t'-\mathbb{E}(\epsilon_t^2\boldsymbol{x}_t\boldsymbol{x}_t')] \\ &\quad=\frac{1}{T}\sum_{t=1}^{T}\left(\epsilon_t^2\boldsymbol{x}_t\boldsymbol{x}_t'-\mathbb{E}(\epsilon_t^2\boldsymbol{x}_t\boldsymbol{x}_t')\right)-\frac{2}{T}\sum_{t=1}^{T}\left(\epsilon_t\boldsymbol{x}_t'(\hat{\boldsymbol{\beta}}_T-\boldsymbol{\beta}_o)\boldsymbol{x}_t\boldsymbol{x}_t'\right)+ \\ &\qquad\frac{1}{T}\sum_{t=1}^{T}\left((\hat{\beta}_T-\boldsymbol{\beta}_o)'\boldsymbol{x}_t\boldsymbol{x}_t'(\hat{\beta}_T-\boldsymbol{\beta}_o)\boldsymbol{x}_t\boldsymbol{x}_t'\right).\end{aligned}$$

The first term on the right-hand side would converge to zero in probability if $\{\epsilon_t^2 \boldsymbol{x}_t \boldsymbol{x}_t'\}$ obeys a WLLN. By noting that under [B2],

$$\mathbb{E}(\epsilon_t \boldsymbol{x}_t' \boldsymbol{x}_t \boldsymbol{x}_t') = \mathbb{E}[\mathbb{E}(\epsilon_t | \mathcal{Y}^{t-1}, \mathcal{W}^t) \boldsymbol{x}_t' \boldsymbol{x}_t \boldsymbol{x}_t'] = \boldsymbol{0},$$

a suitable WLLN will ensure

$$\frac{1}{T} \sum_{t=1}^{T} \epsilon_t \boldsymbol{x}_t' \boldsymbol{x}_t \boldsymbol{x}_t' \xrightarrow{\mathbb{P}} \boldsymbol{0}.$$

This, together with the fact that $\hat{\beta}_T - \boldsymbol{\beta}_o$ is $O_{\mathbb{P}}(T^{-1/2})$, shows that the second term also converges to zero in probability. Similarly, the third term also vanishes in the limit by a suitable WLLN. These results together indicate that, as long as data have proper WLLN effects,

$$\frac{1}{T} \sum_{t=1}^{T} [\hat{e}_t^2 \boldsymbol{x}_t \boldsymbol{x}_t' - \mathbb{E}(\epsilon_t^2 \boldsymbol{x}_t \boldsymbol{x}_t')] \xrightarrow{\mathbb{P}} \boldsymbol{0}.$$

A consistent estimator of $\boldsymbol{V}_o$ is therefore

$$\hat{\boldsymbol{V}}_T = \frac{1}{T} \sum_{t=1}^{T} \hat{e}_t^2 \boldsymbol{x}_t \boldsymbol{x}_t'. \tag{6.11}$$

Thus, $\boldsymbol{V}_o$ can be consistently estimated without modeling the conditional variance $\mathbb{E}(\epsilon_t^2 | \mathcal{Y}^{t-1}, \mathcal{W}^t)$. An estimator of this form is known as a *heteroskedasticity-consistent covariance matrix estimator* which is consistent when conditional heteroskedasticity is present and of unknown form. Consequently, a consistent estimator of $\boldsymbol{D}_o$ is

$$\hat{\boldsymbol{D}}_T = \left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} \hat{e}_t^2 \boldsymbol{x}_t \boldsymbol{x}_t' \right) \left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{x}_t \boldsymbol{x}_t' \right)^{-1}. \tag{6.12}$$

This estimator was proposed by Eicker (1967) and White (1980) and known as the Eicker-White covariance matrix estimator. While the estimator (6.10) is inconsistent under conditionally heteroskedasticity, the Eicker-White estimator is "robust" in the sense that it remains consistent under conditional homoskedasticity and heteroskedasticity. Yet the Eicker-White estimator is less efficient than (6.10) when ϵ_t are in fact conditionally homoskedastic. That is, we obtain a more robust estimator at the expense of (possible) efficiency loss.

6.3.2 When Serial Correlations Are Present

When $\{\boldsymbol{x}_t \epsilon_t\}$ exhibit serial correlations, it is still possible to estimate (6.6) and (6.7) consistently. Let $m(T)$ denote a function of T which diverges to infinity with T but

at a slower rate. Suppose that the autocorrelations between $\boldsymbol{x}_t\epsilon_t$ and $\boldsymbol{x}_{t-\tau}\epsilon_{t-\tau}$ decay sufficiently fast such that

$$\frac{1}{T}\sum_{\tau=m(T)+1}^{T-1}\sum_{t=\tau+1}^{T}\mathbb{E}(\boldsymbol{x}_t\epsilon_t\epsilon_{t-\tau}\boldsymbol{x}_{t-\tau}')\to\boldsymbol{0}.$$

That is, $\boldsymbol{x}_t\epsilon_t$ and $\boldsymbol{x}_{t-\tau}\epsilon_{t-\tau}$ are asymptotically uncorrelated in a proper way. Then for large T, $\boldsymbol{V}_o$ can be well approximated by

$$\boldsymbol{V}_T^*=\frac{1}{T}\sum_{t=1}^{T}\mathrm{var}(\boldsymbol{x}_t\epsilon_t)+\frac{1}{T}\sum_{\tau=1}^{m(T)}\sum_{t=\tau+1}^{T}\mathbb{E}(\boldsymbol{x}_{t-\tau}\epsilon_{t-\tau}\epsilon_t\boldsymbol{x}_t')+\mathbb{E}(\boldsymbol{x}_t\epsilon_t\epsilon_{t-\tau}\boldsymbol{x}_{t-\tau}').$$

Estimating $\boldsymbol{V}_o$ now amounts to estimating $\boldsymbol{V}_T^*$.

White (1984) notes that a consistent estimator of $\boldsymbol{V}_T^*$ is its sample counterpart:

$$\check{\boldsymbol{V}}_T=\frac{1}{T}\sum_{t=1}^{T}\hat{e}_t^2\boldsymbol{x}_t\boldsymbol{x}_t'+\frac{1}{T}\sum_{\tau=1}^{m(T)}\sum_{t=\tau+1}^{T}\left(\boldsymbol{x}_{t-\tau}\hat{e}_{t-\tau}\hat{e}_t\boldsymbol{x}_t'+\boldsymbol{x}_t\hat{e}_t\hat{e}_{t-\tau}\boldsymbol{x}_{t-\tau}'\right),$$

A major problem with this estimator is that $\check{\boldsymbol{V}}_T$ need not be positive semi-definite and hence cannot be a well-defined variance-covariance matrix. Newey and West (1987) show that with a suitable weighting function $w_{m(T)}(\tau)$, the estimator below is guaranteed to be positive semi-definite while remaining consistent for $\boldsymbol{V}_T^*$:

$$\begin{aligned}\hat{\boldsymbol{V}}_T=&\frac{1}{T}\sum_{t=1}^{T}\hat{e}_t^2\boldsymbol{x}_t\boldsymbol{x}_t'+\\&\frac{1}{T}\sum_{\tau=1}^{T-1}w_{m(T)}(\tau)\sum_{t=\tau+1}^{T}\left(\boldsymbol{x}_{t-\tau}\hat{e}_{t-\tau}\hat{e}_t\boldsymbol{x}_t'+\boldsymbol{x}_t\hat{e}_t\hat{e}_{t-\tau}\boldsymbol{x}_{t-\tau}'\right)\end{aligned}\tag{6.13}$$

The estimator (6.13) is known as a *heteroskedasticity and autocorrelation-consistent (HAC) covariance matrix estimator* and is valid when both conditional heteroskedasticity and serial correlations are present but of an unknown form. The resulting consistent estimator of $\boldsymbol{D}_o$ is

$$\hat{\boldsymbol{D}}_T=\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\hat{\boldsymbol{V}}_T\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1},\tag{6.14}$$

where $\hat{\boldsymbol{V}}_T$ is given by (6.13); cf. the Eicker-White estimator (6.12). An estimator of this type is usually referred to as the Newey-West covariance matrix estimator.

In particular, Newey and West (1987) suggested the so-called *Bartlett kernel* for the weighting function of $\hat{\boldsymbol{V}}_T$:

$$w_{m(T)}(\tau) = \begin{cases} 1 - \frac{\tau}{m(T)}, & \text{if } 0 \le \frac{\tau}{m(T)} \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

Gallant (1987) chose the Parzen kernel as the weighting function:

$$w_{m(T)}(\tau) = \begin{cases} 1 - 6\Big(\frac{\tau}{m(T)}\Big)^2 + 6\Big(\frac{\tau}{m(T)}\Big)^3, & \text{if } 0 \le \frac{\tau}{m(T)} \le 1/2, \\ 2\Big(1 - \frac{\tau}{m(T)}\Big)^3, & \text{if } 1/2 \le \frac{\tau}{m(T)} \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

Consider the Bartlett kernel where $w_{m(T)}(\tau) = 1 - \tau/m(T)$. For a fixed $m(T)$, it is decreasing in τ; hence a smaller weight is assigned when two random variables are separated for a long time period (i.e., τ is large). On the other hand, for a fixed τ, $w_{m(T)}(\tau) \to 1$ as $m(T) \to \infty$ and hence entails little loss asymptotically. In practice, a finite number of $m(T)$ must be chosen to compute $\hat{\boldsymbol{V}}_T$. It is worth noting that a small $m(T)$ may result in substantial finite-sample bias. For other choices of weighting functions and a method of determining the approximation lags $m(T)$, we refer to Andrews (1991).

Comparing to the Eicker-White estimator, the Newey-West estimator is robust to both conditional heteroskedasticity of ϵ_t and and serial correlations of $\boldsymbol{x}_t\epsilon_t$, yet the latter is less efficient than the former when $\boldsymbol{x}_t\epsilon_t$ are not serially correlated. Finally, we note that both the Eicker-White estimator (6.12) and the Newey-West estimator (6.14) are non-parametric in nature because they do not rely on any parametric model of conditional heteroskedasticity and serial correlations.

6.4 Large-Sample Tests

After learning the asymptotic properties of the OLS estimator under more general conditions, it is now important to construct suitable tests and derive their limiting distributions. In this section, we will study two large-sample tests for the linear hypothesis

$$H_0\colon\ \boldsymbol{R}\boldsymbol{\beta}^* = \boldsymbol{r},$$

where $\boldsymbol{R}$ is a $q \times k$ ($q < k$) nonstochastic matrix with rank q, and $\boldsymbol{r}$ is a pre-specified real vector, as in Section 3.4.

6.4.1 Wald Test

Given that the OLS estimator $\hat{\boldsymbol{\beta}}_T$ is consistent for some parameter vector $\boldsymbol{\beta}^*$, one would expect that $\boldsymbol{R}\hat{\boldsymbol{\beta}}_T$ is "close" to $\boldsymbol{R}\boldsymbol{\beta}^*$ when T becomes large. As $\boldsymbol{R}\boldsymbol{\beta}^* = \boldsymbol{r}$ under the null hypothesis, whether $\boldsymbol{R}\hat{\boldsymbol{\beta}}_T$ is sufficiently "close" to $\boldsymbol{r}$ constitutes an evidence for or against the null hypothesis. The *Wald test* is based on the difference between $\boldsymbol{R}\hat{\boldsymbol{\beta}}_T$ and $\boldsymbol{r}$.

When [B1] and [B3] hold, we have from Theorem 6.9 that

$$\sqrt{T}\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{R}\boldsymbol{D}_o\boldsymbol{R}'),$$

or equivalently,

$$\boldsymbol{\Gamma}_o^{-1/2}\sqrt{T}\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_q),$$

where $\boldsymbol{\Gamma}_o = \boldsymbol{R}\boldsymbol{D}_o\boldsymbol{R}' = \boldsymbol{R}\boldsymbol{M}_{xx}^{-1}\boldsymbol{V}_o\boldsymbol{M}_{xx}^{-1}\boldsymbol{R}'$. By Theorem 6.10, asymptotic normality would not be affected if $\boldsymbol{\Gamma}_o$ is replaced by its consistent estimator, e.g.,

$$\hat{\boldsymbol{\Gamma}}_T = \boldsymbol{R}\hat{\boldsymbol{D}}_T\boldsymbol{R}' = \boldsymbol{R}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\hat{\boldsymbol{V}}_T\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\boldsymbol{R}',$$

where $\hat{\boldsymbol{V}}_T$ is a consistent estimator of $\boldsymbol{V}_o$. That is,

$$\hat{\boldsymbol{\Gamma}}_T^{-1/2}\sqrt{T}\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_q). \tag{6.15}$$

Under the null hypothesis, $\boldsymbol{R}\boldsymbol{\beta}^* = \boldsymbol{r}$, the Wald test statistic is the inner product of (6.15):

$$\mathcal{W}_T = T(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})'\hat{\boldsymbol{\Gamma}}_T^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}). \tag{6.16}$$

The result below follows directly from the continuous mapping theorem (Lemma 5.20).

Theorem 6.13 *Given the linear specification* (6.1)*, suppose that* [B1] *and* [B3] *hold. Then, under the null hypothesis,*

$$\mathcal{W}_T \xrightarrow{D} \chi^2(q).$$

where $\mathcal{W}_T$ is given by (6.16) *and q is the number of hypotheses.*

The Wald test has much wider applicability because it is valid for a wide variety of data which may be non-Gaussian, heteroskedastic, and serially correlated. What matter here are the asymptotic normality result of the OLS estimator and a consistent estimator of $\boldsymbol{V}_o$. If an inconsistent estimator of $\boldsymbol{V}_o$ is used in the test statistic, both $\hat{\boldsymbol{D}}_T$ and $\hat{\boldsymbol{\Gamma}}_T$ become inconsistent, and, consequently, the Wald statistic $\mathcal{W}_T$ will not have a limiting χ^2 distribution.

Example 6.14 Test of a subset of coefficients being zero: Given the linear specification

$$y_t = \boldsymbol{x}_{1,t}\boldsymbol{b}_1 + \boldsymbol{x}_{2,t}\boldsymbol{b}_2 + e_t,$$

where $\boldsymbol{x}_{1,t}$ is $(k-s)\times 1$ and $\boldsymbol{x}_{2,t}$ is $s\times 1$, suppose that the specification is correct for the conditional mean with $\boldsymbol{\beta}_o = [\boldsymbol{b}_{1,o}' \; \boldsymbol{b}_{2,o}']'$. If we want to verify whether

$$\mathbb{E}(y_t|\mathcal{Y}^{t-1}, \mathcal{W}^t) = \boldsymbol{x}_{1,t}\boldsymbol{b}_{1,o},$$

then the hypothesis is $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{0}$ with $\boldsymbol{R} = [\boldsymbol{0}_{s\times(k-s)} \; \boldsymbol{I}_s]$. The Wald test statistic is

$$\mathcal{W}_T = T\hat{\boldsymbol{\beta}}_T'\boldsymbol{R}'\Big[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\hat{\boldsymbol{V}}_T(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}'\Big]^{-1}\boldsymbol{R}\hat{\boldsymbol{\beta}}_T \xrightarrow{D} \chi^2(s).$$

Note that this is a general expression of the Wald statistic; its exact form depends on $\hat{\boldsymbol{V}}_T$.

When $\boldsymbol{V}_o$ can be consistently estimated by $\hat{\boldsymbol{V}}_T = \sum_{t=1}^T \hat{e}_t^2\boldsymbol{x}_t\boldsymbol{x}_t'/T$,

$$\mathcal{W}_T = \hat{\boldsymbol{\beta}}_T'\boldsymbol{R}'\left[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{t=1}^T \hat{e}_t^2\boldsymbol{x}_t\boldsymbol{x}_t'\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\right]^{-1}\boldsymbol{R}\hat{\boldsymbol{\beta}}_T.$$

If the null hypothesis is that the i^{th} coefficient is zero, then $\boldsymbol{R}$ is the i^{th} Cartesian unit vector $\boldsymbol{c}_i$ so that

$$\hat{\beta}_{i,T}/\sqrt{\hat{d}_{ii}} \xrightarrow{D} N(0,1), \tag{6.17}$$

where $\hat{d}_{ii}$ is the i^{th} diagonal element of $\hat{D}_T/T$:

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\left(\sum_{t=1}^T \hat{e}_t^2\boldsymbol{x}_t\boldsymbol{x}_t'\right)(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Note that $\hat{d}_{ii}$ is usually referred to as White's estimate of the standard error of the i^{th} coefficient. In view of (6.17), we can test the significance of the i^{th} coefficient using the t statistic with the OLS standard error replaced by the Eicker-White estimate of the standard error. We can also base other t tests on the Eicker-White standard errors.

When a consistent estimator of $\boldsymbol{V}_o$ is $\hat{\boldsymbol{V}}_T = \hat{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)$, the Wald statistic becomes

$$\mathcal{W}_T = \hat{\boldsymbol{\beta}}_T'\boldsymbol{R}'\Big[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\Big]^{-1}\boldsymbol{R}\hat{\boldsymbol{\beta}}_T/\hat{\sigma}_T^2,$$

which is nothing but s times the standard F-statistic. Therefore, for testing the i^{th} coefficient being zero, the standard t ratio will still do. The difference is that, instead of having an exact null distribution, the t ratio now has the asymptotic null distribution $N(0,1)$. □

Remark: The F-test-based version of the Wald test is appropriate only when $\hat{\boldsymbol{V}}_T = \hat{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)$ is consistent for $\boldsymbol{V}_o$. We know that if ϵ_t are conditionally heteroskedastic and/or $\boldsymbol{x}_t\epsilon_t$ are serially correlated, this estimator is not consistent for $\boldsymbol{V}_o$. Consequently, the F-test-based version does not have a limiting χ^2 distribution.

6.4.2 Lagrange Multiplier Test

From Section 3.4.3 we have seen that, given the constraint $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$, the constrained OLS estimator can be obtained by finding the saddle point of the Lagrangian:

$$\frac{1}{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{r})'\boldsymbol{\lambda},$$

where $\boldsymbol{\lambda}$ is the $q \times 1$ vector of Lagrange multipliers. The underlying idea of the Lagrange Multiplier (LM) test of this constraint is to check whether $\boldsymbol{\lambda}$ is sufficiently "close" to zero. Intuitively, $\boldsymbol{\lambda}$ can be interpreted as the "shadow price" of this constraint and hence should be "small" when the constraint is valid (i.e., the null hypothesis is true); otherwise, $\boldsymbol{\lambda}$ ought to be "large." Again, the closeness between $\boldsymbol{\lambda}$ and zero must be determined by the distribution of the estimator of $\boldsymbol{\lambda}$.

It is easy to find the solutions to the Lagrangian above:

$$\ddot{\boldsymbol{\lambda}}_T = 2[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}),$$
$$\ddot{\boldsymbol{\beta}}_T = \hat{\boldsymbol{\beta}}_T - (\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}'\ddot{\boldsymbol{\lambda}}_T/2.$$

Here, $\ddot{\boldsymbol{\beta}}_T$ is the constrained OLS estimator of $\boldsymbol{\beta}$, and $\ddot{\boldsymbol{\lambda}}_T$ is the basic ingredient of the LM test. Let $\epsilon_t = y_t - \boldsymbol{x}_t'\boldsymbol{\beta}^*$, where $\boldsymbol{\beta}^*$ satisfies the constraint $\boldsymbol{R}\boldsymbol{\beta}^* = \boldsymbol{r}$ under the null hypothesis and

$$\boldsymbol{V}_o = \lim_{T\to\infty} \operatorname{var}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{x}_t\epsilon_t\right).$$

Given [B1] and [B3],

$$\begin{aligned}\sqrt{T}\ddot{\boldsymbol{\lambda}}_T &= 2\left[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}'\right]^{-1}\sqrt{T}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}) \\ &\xrightarrow{D} 2(\boldsymbol{R}\boldsymbol{M}_{xx}^{-1}\boldsymbol{R}')^{-1}\, N(\boldsymbol{0}, \boldsymbol{R}\boldsymbol{D}_o\boldsymbol{R}').\end{aligned}$$

where $\boldsymbol{D}_o = \boldsymbol{M}_{xx}^{-1}\boldsymbol{V}_o\boldsymbol{M}_{xx}^{-1}$ and the limiting distribution of the right-hand side is $N(\boldsymbol{0}, \boldsymbol{\Lambda}_o)$ with

$$\boldsymbol{\Lambda}_o = 4(\boldsymbol{R}\boldsymbol{M}_{xx}^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\boldsymbol{D}_o\boldsymbol{R}')(\boldsymbol{R}\boldsymbol{M}_{xx}^{-1}\boldsymbol{R}')^{-1}.$$

We immediately have

$$\boldsymbol{\Lambda}_o^{-1/2}\sqrt{T}\ddot{\boldsymbol{\lambda}}_T \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_q).$$

Similar as before, this result remains valid when $\boldsymbol{\Lambda}_o$ is replaced by a consistent estimator; for example,

$$\ddot{\boldsymbol{\Lambda}}_T = 4\Big[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}'\Big]^{-1}\Big[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\ddot{\boldsymbol{V}}_T(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}'\Big] \Big[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}'\Big]^{-1},$$

where $\ddot{\boldsymbol{V}}_T$ is a consistent estimator of $\boldsymbol{V}_o$ based on the constrained estimation result. Typically, $\ddot{\boldsymbol{V}}_T$ involves $\boldsymbol{x}_t$ and constrained OLS residuals $\ddot{e}_t = y_t - \boldsymbol{x}_t'\ddot{\boldsymbol{\beta}}_T$. Thus,

$$\ddot{\boldsymbol{\Lambda}}_T^{-1/2}\sqrt{T}\ddot{\boldsymbol{\lambda}}_T \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_q). \tag{6.18}$$

and the LM statistic is the inner product of (6.18):

$$\mathcal{LM}_T = T\ddot{\boldsymbol{\lambda}}_T'\ddot{\boldsymbol{\Lambda}}_T^{-1}\ddot{\boldsymbol{\lambda}}_T, \tag{6.19}$$

The result below is again a consequence of the continuous mapping theorem.

Theorem 6.15 *Given the linear specification* (6.1)*, suppose that* [B1] *and* [B3] *hold. Then, under the null hypothesis,*

$$\mathcal{LM}_T \xrightarrow{D} \chi^2(q),$$

where $\mathcal{LM}_T$ *is given by* (6.19).

Let $\ddot{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\ddot{\boldsymbol{\beta}}_T$ denote the vector of constrained OLS residuals. By noting that the constrained OLS estimator must satisfy the constraint (i.e., $\boldsymbol{R}\ddot{\boldsymbol{\beta}}_T = \boldsymbol{r}$), we can write

$$\begin{aligned}\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r} &= \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\ddot{\boldsymbol{\beta}}_T)/T \\ &= \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{X}'\ddot{\boldsymbol{e}}/T.\end{aligned}$$

An alternative expression of $\ddot{\boldsymbol{\lambda}}_T$ is then

$$\ddot{\boldsymbol{\lambda}}_T = 2[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}']^{-1}\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{X}'\ddot{\boldsymbol{e}}/T.$$

It follows that the LM test statistic is algebraically equivalent to

$$\mathcal{LM}_T = T\ddot{\boldsymbol{e}}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\Big[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\ddot{\boldsymbol{V}}_T(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}'\Big]^{-1} \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\ddot{\boldsymbol{e}}. \tag{6.20}$$

This expression shows that only constrained estimation is needed to compute the LM statistic.

A fundamental difference between the LM and Wald tests is that the former involves only constrained estimation, whereas the latter requires unconstrained estimation. As such, the Wald test would be more convenient if constrained specifications are difficult to estimate, such as a linear specification with a nonlinear constraint. The LM test, on the other hand, would be simpler if constrained estimation is easier to compute. Similar to the Wald test, the LM test is also valid for a wide variety of data which may be non-Gaussian, heteroskedastic, and serially correlated. Again, the asymptotic normality result of the OLS estimator and consistent estimation of $\boldsymbol{V}_o$ play a crucial role in Theorem 6.15. If an inconsistent estimator of $\boldsymbol{V}_o$ is used to construct $\ddot{\boldsymbol{\Lambda}}_T$, the resulting LM test will not have a limiting χ^2 distribution.

Example 6.16 Test of a subset of coefficients being zero: Given the following specification:

$$y_t = \boldsymbol{x}_{1,t}\boldsymbol{b}_1 + e_t,$$

suppose that we want to know whether

$$\mathbb{E}(y_t|\mathcal{Y}^{t-1}, \mathcal{W}^t) = \boldsymbol{x}_{1,t}\boldsymbol{b}_{1,o} + \boldsymbol{x}_{2,t}\boldsymbol{b}_{2,o},$$

where $\boldsymbol{x}_{1,t}$ is $(k-s)\times 1$ and $\boldsymbol{x}_{2,t}$ is $s\times 1$. The specification would be correct for the conditional mean if $\boldsymbol{b}_{2,o} = \boldsymbol{0}$. Letting $\boldsymbol{\beta}_o = [\boldsymbol{b}_{1,o}' \; \boldsymbol{b}_{2,o}']'$, the null hypothesis is $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{0}$ with $\boldsymbol{R} = [\boldsymbol{0}_{s\times(k-s)} \; \boldsymbol{I}_s]$. The specification above is then a constrained version of

$$y_t = \boldsymbol{x}_{1,t}\boldsymbol{b}_1 + \boldsymbol{x}_{2,t}\boldsymbol{b}_2 + e_t.$$

For this specification, the constrained OLS estimator is $\ddot{\boldsymbol{\beta}}_T = (\ddot{\boldsymbol{b}}_{1,T}' \; \boldsymbol{0}')'$, where

$$\ddot{\boldsymbol{b}}_{1,T} = \left(\sum_{t=1}^T \boldsymbol{x}_{1,t}\boldsymbol{x}_{1,t}'\right)^{-1} \sum_{t=1}^T \boldsymbol{x}_{1,t}y_t = (\boldsymbol{X}_1'\boldsymbol{X}_)^{-1}\boldsymbol{X}_1'\boldsymbol{y},$$

which is obtained from the constrained specification. The LM statistic now can be computed as (6.20) with $\boldsymbol{X} = [\boldsymbol{X}_1 \; \boldsymbol{X}_2]$ and the vector of constrained OLS residuals $\ddot{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}_1\ddot{\boldsymbol{b}}_{1,T}$.

When $\ddot{\boldsymbol{V}}_T = \ddot{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)$ is consistent for $\boldsymbol{V}_o$ under the null hypothesis, where $\ddot{\sigma}_T^2 = \sum_{t=1}^T \ddot{e}_t^2/(T-k+s)$, we have

$$(\boldsymbol{R}\ddot{\boldsymbol{D}}_T\boldsymbol{R}')^{-1} = \frac{1}{\ddot{\sigma}_T^2}\Big[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}\boldsymbol{R}'\Big]^{-1}.$$

It can be verified that, by the Frisch-Waugh-Lovell Theorem,

$$\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}' = [\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2]^{-1},$$

$$\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = [\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2]^{-1}\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{P}_1),$$

where $\boldsymbol{P}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$. The LM statistic now simplifies to

$$\begin{aligned}\mathcal{LM}_T &= \ddot{\boldsymbol{e}}'(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2[\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2]^{-1}\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{P}_1)\ddot{\boldsymbol{e}}/\ddot{\sigma}_T^2 \\ &= \ddot{\boldsymbol{e}}'\boldsymbol{X}_2[\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2]^{-1}\boldsymbol{X}_2'\ddot{\boldsymbol{e}}/\ddot{\sigma}_T^2,\end{aligned}$$

because $\boldsymbol{X}_1'\ddot{\boldsymbol{e}} = \boldsymbol{0}$ so that $\boldsymbol{P}_1\ddot{\boldsymbol{e}} = \boldsymbol{0}$. As $\ddot{\sigma}_T^2 = \ddot{\boldsymbol{e}}'\ddot{\boldsymbol{e}}/(T-k+s)$,

$$\mathcal{LM}_T = \frac{\ddot{\boldsymbol{e}}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\ddot{\boldsymbol{e}}}{\ddot{\boldsymbol{e}}'\ddot{\boldsymbol{e}}/(T-k+s)} = (T-k+s)R^2,$$

where R^2 is the (non-centered) coefficient of determination of regressing $\ddot{\boldsymbol{e}}$ on $\boldsymbol{X}$. If the estimator $\ddot{\sigma}_T^2 = \sum_{t=1}^T \ddot{e}_t^2/T$ is used, we simply have TR^2 as the test statistic. It must be emphasized that the simpler version of the LM statistic is valid only when $\ddot{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)$ is a consistent estimator of $\boldsymbol{V}_o$; otherwise, TR^2 does not have a limiting χ^2 distribution. If the LM statistic is based on the heteroskedasticity-consistent covariance matrix estimator:

$$\ddot{\boldsymbol{V}}_T = \frac{1}{T}\sum_{t=1}^T \ddot{e}_t^2\boldsymbol{x}_t\boldsymbol{x}_t',$$

it cannot be simplified to TR^2.

As the LM test only requires constrained estimation, it is based on the simpler, constrained specification and checks whether additional s regressors should be included as well. Comparing to Example 6.14, the Wald test checks whether the unconstrained specification should exclude s redundant regressors. Thus, the LM test permits testing "up" (from a simplerr specification), while the Wald test can be employed to test "down" (from a more complex specification). □

Remark: It can also be shown that the Wald and LM statistics are asymptotically equivalent under the null hypothesis, i.e.,

$$\mathcal{W}_T - \mathcal{LM}_T \xrightarrow{\mathbb{P}} 0;$$

see Exercise 6.9. If $\boldsymbol{V}_o$ is known, these two statistics turn out to be algebraically equivalent. Note, however, that these two tests may result in conflicting statistical inferences in finite samples. For instance, it can be shown that when there are no heteroskedasticity and serial correlations, $\mathcal{W}_T \geq \mathcal{LM}_T$ in numerical values; see e.g., Godfrey (1988) for more details.

6.5 Application: Autoregressive Models

To analyze time series data, it is quite common to postulate an *autoregressive* (AR) specification, in the sense that the regressors are nothing but the lagged dependent variables. In particular, an AR(p) specification is such that

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + e_t, \qquad t = p+1, \ldots, T.$$

The specification in Examples 6.6 and 6.8 is AR(1) without the constant term. The OLS estimators of $\beta_0, \ldots, \beta_p$ are obtained by regressing y_t on $\boldsymbol{\eta}_{t-1} = [y_{t-1} \ \cdots \ y_{t-p}]'$ for $t = p+1, \ldots, T$. The OLS variance estimator is

$$\hat{\sigma}_T^2 = \frac{1}{T-2p-1} \sum_{t=p+1}^{T} \hat{e}_t^2,$$

where $\hat{e}_t^2$ are the OLS residuals. It is also common to compute the variance estimator as the sum of squared residuals divided by T or $T-p$. The properties of the OLS estimators depend crucially on whether y_t are weakly stationary.

6.5.1 Properties of the OLS estimators

Recall that $\{y_t\}$ is weakly stationary if its mean, variance and autocovariances are all independent of t. When $\{y_t\}$ is weakly stationary with finite fourth moment, both y_t, $y_t \boldsymbol{\eta}_{t-1}$ and $\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}'$ obey a WLLN.

Let $\mu_o = \mathbb{E}(y_t)$ and $\gamma_j = \operatorname{cov}(y_t, y_{t-j})$ for $j = 0, \pm 1, \pm 2, \ldots$. Clearly, γ_0 is the variance of y_t and $\gamma_j = \gamma_{-j}$. Then,

$$\boldsymbol{\Gamma} = \operatorname{var}(\boldsymbol{\eta}_{t-1}) = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{p-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \gamma_{p-3} & \cdots & \gamma_0 \end{bmatrix}.$$

The WLLN effect ensures:

$$\frac{1}{T} \sum_{t=p+1}^{T} y_t \xrightarrow{\mathbb{P}} \mu_o,$$

$$\frac{1}{T} \sum_{t=p+1}^{T} y_t \boldsymbol{\eta}_{t-1} \xrightarrow{\mathbb{P}} [\gamma_1 + \mu_o^2 \ \cdots \ \gamma_p + \mu_o^2]',$$

$$\frac{1}{T} \sum_{t=p+1}^{T} \boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}' \xrightarrow{\mathbb{P}} \boldsymbol{\Gamma} + \mu_o^2 \boldsymbol{\ell}\boldsymbol{\ell}',$$

where $\boldsymbol{\ell}$ is the $p \times 1$ vector of ones. It follows that

$$\hat{\boldsymbol{\beta}}_T \xrightarrow{\mathbb{P}} \begin{bmatrix} 1 & \mu_o \boldsymbol{\ell}' \\ \mu_o \boldsymbol{\ell} & \boldsymbol{\Gamma} + \mu_o^2 \boldsymbol{\ell}\boldsymbol{\ell}' \end{bmatrix}^{-1} \begin{bmatrix} \mu_o \\ \gamma_1 + \mu_o^2 \\ \vdots \\ \gamma_p + \mu_o^2 \end{bmatrix}.$$

In particular, if the specification does not contain the constant term and $\boldsymbol{\eta}_{t-1}$ contains only y_{t-1}, the OLS estimator converges in probability to

$$(\gamma_1 + \mu_o^2)/(\gamma_0 + \mu_o^2).$$

When $\mu_o = 0$, this probability limit simplifies to γ_1/γ_0, which is precisely what we obtained in Example 6.2.

If y_t is generated as an AR(p) process:

$$y_t = c_o + \psi_1 y_{t-1} + \psi_2 y_{t-2} + \cdots + \psi_p y_{t-p} + \epsilon_t,$$

the true parameters $c, \psi_1, \ldots, \psi_p$ must satisfy certain constraints so as to ensure weak stationarity (see below). In addition, if

$$\mathbb{E}(y_t | \mathcal{Y}^{t-1}) = c_o + \psi_1 y_{t-1} + \psi_2 y_{t-2} + \cdots + \psi_p y_{t-p},$$

Theorem 6.3 ensures that the OLS estimators will converge in probability to the true parameters. Note, however, that $\{\epsilon_t\}$ may be a white noise but not a martingale difference sequence.

Whether the AR(p) specification is a correct specification for the conditional mean function, the resulting OLS estimators, with suitable normalization, are asymptotically normally distributed.

6.5.2 Difference Equation

Suppose that y_t are generated according to the following first-order difference equation:

$$y_t = \psi_1 y_{t-1} + u_t, \qquad t = 0, 1, 2, \ldots .$$

It is easily verified that, by recursive substitution,

$$y_{t+j} = \psi_1^{j+1} y_{t-1} + \psi_1^j u_t + \psi_1^{j-1} u_{t+1} + \cdots + \psi_1 u_{t+j-1} + u_{t+j}.$$

Define the *dynamic multiplier* of u_t on y_{t+j} as

$$\partial y_{t+j} / \partial u_t = \psi_1^j,$$

which is the effect of a given change of u_t on y_{t+j}. When $|\psi_1| < 1$, the dynamic multiplier approaches zero when j tends to infinity so that the effect of u_t eventually dies out. As y_t does not depend much on what what happens in the distant past, the difference equation is said to be *stable*. When $|\psi_1| > 1$, the difference equation is *explosive* in the sense that the effect of u_t on future y's grows exponentially fast. If $\psi_1 = 1$, u_t has a constant effect on future y's.

Consider now a p th-order difference equation:

$$y_t = \psi_1 y_{t-1} + \psi_2 y_{t-2} + \cdots + \psi_p y_{t-p} + u_t,$$

which can be expressed as a first-order vector difference equation:

$$\boldsymbol{\eta}_t = \boldsymbol{F}\boldsymbol{\eta}_{t-1} + \boldsymbol{\nu}_t,$$

with

$$\boldsymbol{\eta}_t = \begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p+1} \end{bmatrix}, \qquad \boldsymbol{F} = \begin{bmatrix} \psi_1 & \psi_2 & \cdots & \psi_{p-1} & \psi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \qquad \boldsymbol{\nu}_t = \begin{bmatrix} u_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Recursive substitution yields

$$\boldsymbol{\eta}_{t+j} = \boldsymbol{F}^{j+1}\boldsymbol{\eta}_{t-1} + \boldsymbol{F}^{j}\boldsymbol{\nu}_t + \boldsymbol{F}^{j-1}\boldsymbol{\nu}_{t+1} + \cdots + \boldsymbol{F}\boldsymbol{\nu}_{t+j-1} + \boldsymbol{\nu}_{t+j}.$$

The dynamic multiplier of $\boldsymbol{\nu}_t$ on $\boldsymbol{\eta}_{t+j}$ is

$$\nabla_{\boldsymbol{\nu}_t}\boldsymbol{\eta}_{t+j} = \boldsymbol{F}^j,$$

and its (m,n) th element is denoted as $f^j(m,n)$. It is straightforward to verify that

$$y_{t+j} = f^{j+1}(1,1)y_{t-1} + \cdots + f^{j+1}(1,p)y_{t-p} + \sum_{i=0}^{j} f^i(1,1)u_{t+j-i}.$$

The dynamic multiplier of u_t on y_{t+j} is thus

$$\partial y_{t+j}/\partial u_t = f^j(1,1),$$

the $(1,1)$ th element of $\boldsymbol{F}^j$.

Recall that the eigenvalues of $\boldsymbol{F}$ solve the equation: $\det(\boldsymbol{F} - \lambda\boldsymbol{I}) = 0$, which is known as the em characteristic equation of $\boldsymbol{F}$. This equation is of the following form:

$$\lambda^p - \psi_1\lambda^{p-1} - \cdots - \psi_{p-1}\lambda - \psi_p = 0.$$

When all the eigenvalues of $\boldsymbol{F}$ are distinct, then $\boldsymbol{F}$ can be diagonalized by a nonsingular matrix $\boldsymbol{C}$ such that $\boldsymbol{C}^{-1}\boldsymbol{F}\boldsymbol{C} = \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is the diagonal matrix with all the eigenvalues of $\boldsymbol{F}$ on its main diagonal. Writing $\boldsymbol{F} = \boldsymbol{C}\boldsymbol{\Lambda}\boldsymbol{C}^{-1}$, we have $\boldsymbol{F}^j = \boldsymbol{C}\boldsymbol{\Lambda}^j\boldsymbol{C}^{-1}$. When all the eigenvalues of $\boldsymbol{F}$ are less than one in modulus (i.e., for any complex eigenvalue $\lambda^* = a + bi$, $i = (-1)^{1/2}$, $|\lambda^*| = (a^2 + b^2)^{1/2} < 1$), $\boldsymbol{\Lambda}^j$ tends to the zero matrix as j goes to infinity, and so does $\boldsymbol{F}^j$. In this case, f_{11}^j will be approaching zero, so that the difference equation is stable. Thus, a p th-order difference equation is stable provided that the eigenvalues of $\boldsymbol{F}$ (the roots of the characteristic equation) are all less than one in modulus. This is equivalent to saying that these roots must lie inside the unit circle on the complex plane. This condition requires that the coefficients ψ_i must satisfy certain constraints.

If there is a root of the characteristic equation equals one in modulus (i.e., on the unit circle), such a root is usually referred to as a *unit root*. When the characteristic equation has a unit root with all remaining roots less than one in modulus,

$$\lim_{j\to\infty} \boldsymbol{F}^j = \boldsymbol{C} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \boldsymbol{C}^{-1},$$

so that its $(1,1)$ th element $f^j(1,1)$ is a constant. If there is at least one eigenvalue greater than one in modulus, this eigenvalue eventually dominates, and $\boldsymbol{F}^j$ will explode. The difference equation is then explosive.

Alternatively, setting $\lambda = 1/z$ and multiplying the characteristic equation by z^p we obtain:

$$1 - \psi_1 z - \cdots - \psi_{p-1} z^{p-1} - \psi_p z^p = 0.$$

The condition that all the roots of the characteristic equation are inside the unit circle is thus equivalent to requiring all the roots of the polynomial above being outside the unit circle.

6.5.3 Weak Stationarity

Let $\mathfrak{B}$ denote the *back-shift operator* in the sense that $\mathfrak{B}y_t = y_{t-1}$. We will also write $\mathfrak{B}^2 y_t = \mathfrak{B}(\mathfrak{B}y_t) = y_{t-2}$, $\mathfrak{B}^3 y_t = \mathfrak{B}(\mathfrak{B}^2 y_t) = y_{t-3}$, and so on. This operator has no effect on constants, i.e., for any constant c, $\mathfrak{B}c = c$. Hence, the back-shift operator has the

linear property:

$$\mathfrak{B}(cy_t + dz_t) = c(\mathfrak{B}y_t) + d(\mathfrak{B}z_t) = cy_{t-1} + dz_{t-1},$$

for any constants c and d.

Consider now the AR(p) process

$$y_t = c + \psi_1 y_{t-1} + \cdots + \psi_p y_{t-p} + \epsilon_t,$$

which can be written compactly as

$$\Psi(\mathfrak{B})y_t = c + \epsilon_t,$$

where $\Psi(\mathfrak{B}) = 1 + \psi_1\mathfrak{B} + \cdots + \psi_p\mathfrak{B}^p$ is a polynomial in $\mathfrak{B}$. When $\Psi(\mathfrak{B}) = 1 - \psi_1\mathfrak{B}$, it is an AR(1) process. As discussed above, this system would be stable if all the roots of $\Psi(z) = 0$ are outside the unit circle.

Consider again an AR(1) process:

$$(1 - \psi_1\mathfrak{B})y_t = c + \epsilon_t.$$

Note that for large t,

$$(1 + \psi_1\mathfrak{B} + \psi_1^2\mathfrak{B}^2 + \cdots + \psi_1^t\mathfrak{B}^t)(1 - \psi_1\mathfrak{B})y_t = (1 - \psi_1^{t+1}\mathfrak{B}^{t+1})y_t \approx y_t,$$

provided that $\mathfrak{B}^{t+1}y_t = y_{-1}$ is finite and $|\psi_1| < 1$. This suggests that, when $|\psi_1| < 1$, the inverse of $(1 - \psi_1\mathfrak{B})$ can be defined as

$$(1 - \psi_1\mathfrak{B})^{-1} = \lim_{t\to\infty}(1 + \psi_1\mathfrak{B} + \psi_1^2\mathfrak{B}^2 + \cdots + \psi_1^t\mathfrak{B}^t),$$

so that $(1 - \psi_1\mathfrak{B})^{-1}(1 - \psi_1\mathfrak{B}) = \mathfrak{I}$, the identity operator. It follows that

$$y_t = (1 - \psi_1\mathfrak{B})^{-1}(c + \epsilon_t) = \frac{c}{1 - \psi_1} + (1 - \psi_1\mathfrak{B})^{-1}\epsilon_t.$$

When $\{\epsilon_t\}$ is a white noise with mean zero and variance σ_ϵ^2, we find $\mathbb{E}(y_t) = c/(1 - \psi_1)$ and

$$\begin{aligned}
\gamma_0 &= (1 + \psi_1^2 + \psi_1^4 + \cdots)\sigma_\epsilon^2 = \sigma_\epsilon^2/(1 - \psi_1^2),\\
\gamma_1 &= (\psi_1 + \psi_1^3 + \psi_1^5 + \cdots)\sigma_\epsilon^2 = \psi_1[\sigma_\epsilon^2/(1 - \psi_1^2)],\\
\gamma_2 &= (\psi_1^2 + \psi_1^4 + \psi_1^6 + \cdots)\sigma_\epsilon^2 = \psi_1^2[\sigma_\epsilon^2/(1 - \psi_1^2)],\\
&\vdots\\
\gamma_j &= \psi_1^j\frac{\sigma_\epsilon^2}{1 - \psi_1^2}.
\end{aligned}$$

Thus, y_t have a constant mean, constant variance, and autocovariances depending on j but not on t. This shows that $\{y_t\}$ is a weakly stationary process. On the other hand, y_t cannot be weakly stationary when the difference equation is not stable. Note also that the autocovariances can be expressed as

$$\gamma_j = \psi_1^{j-1}\gamma_{j-1} = \psi_1^j\gamma_0, \qquad j = 0, 1, 2, \ldots ,$$

and the autocorrelations are $\rho_j = \psi_1^j = \psi_1\rho_{j-1}$. That is, both the autocovariances and autocorrelations have the same AR(1) structure. If we view the autocorrelations of a process as its "memory," a weakly stationary AR(1) process has exponentially decaying memory and is also said to be of "short memory."

The previous results are readily generalized. For the AR(p) processes $\Psi(\mathfrak{B})y_t = c + \epsilon_t$, where $\Psi(\mathfrak{B})$ is a p^{th}-order polynomial in $\mathfrak{B}$. When all the roots of $\Psi(z)$ are outside the unit circle, y_t are weakly stationary with $\mathbb{E}(y_t) = c/(1-\psi_1-\psi_2-\cdots-\psi_p)$, autocovariances: acut

$$\gamma_0 = \psi_1\gamma_1 + \psi_2\gamma_2 + \cdots + \psi_p\gamma_p + \sigma_\epsilon^2,$$
$$\gamma_j = \psi_1\gamma_{j-1} + \psi_2\gamma_{j-2} + \cdots + \psi_p\gamma_{j-p}, \qquad j = 1, 2, \ldots ,$$

and autocorrelations:

$$\rho_j = \psi_1\rho_{j-1} + \psi_2\rho_{j-2} + \cdots + \psi_p\rho_{j-p}, \qquad j = 1, 2, \ldots .$$

The equation for autocorrelations is also known as the *Yule-Walker equation* which has the same AR(p) structure. As the initial value $\rho_0 = 1$, it is then clear that $\rho_j \to 0$ exponentially fast as j tends to infinity. Hence, a weakly stationary AR(p) process is also a "short memory" process.

6.6 Limitations of the Linear Specification

In this chapter the classical conditions are relaxed so as to allow for more general data in linear regressions. Careful readers must have noticed that, aside from the conditions on the stochastic properties of data, there is always a condition of correct specification ([A2](i) in Chapter 3 and [B2] in this chapter). Such a condition may be too strong in practice, as discussed in Section 3.7. In the context of this chapter, we also notice that, while $\mathbb{E}(y_t|\mathcal{Y}^{t-1}, \mathcal{W}^t)$ must be a function of the elements of $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$, [B2] requires this function being linear. A sufficient condition for linear conditional mean function is that all the elements of $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$ are jointly normally distributed; this condition

is also much too strong in practic, however. If joint normality is unlikely, there would be no guarantee that [B2] is true. Hence, the OLS estimator may converge to some parameter vector that does not have any meaningful interpretations. This suggests that we should not confine ourselves to linear specifications and may want to explore nonlinear specifications instead. The least squares theory for nonlinear specifications is the topic to which we now turn.

Exercises

6.1 Suppose that $y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_o + \epsilon_t$ such that $\mathbb{E}(\epsilon_t) = 0$ for all t.

(a) If $\{\boldsymbol{x}_t\}$ and $\{\epsilon_t\}$ are two mutually independent sequences, i.e., $\boldsymbol{x}_t$ and ϵ_τ are independent for any t and τ, is $\hat{\boldsymbol{\beta}}_T$ unbiased?

(b) If $\{\boldsymbol{x}_t\}$ and $\{\epsilon_t\}$ are two mutually uncorrelated sequences, i.e., $\mathbb{E}(\boldsymbol{x}_t\epsilon_\tau) = \mathbf{0}$ for any t and τ, is $\hat{\boldsymbol{\beta}}_T$ unbiased?

6.2 Consider the specification $y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t$, where $\boldsymbol{x}_t$ is $k \times 1$. Suppose that

$$\mathbb{E}(y_t|\mathcal{Y}^{t-1}, \mathcal{W}^t) = \boldsymbol{z}_t'\boldsymbol{\gamma}_o,$$

where $\boldsymbol{z}_t$ is $m \times 1$. Assuming suitable weak laws for $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$, what is the probability limit of the OLS estimator of $\boldsymbol{\beta}$?

6.3 Consider the specification $y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \boldsymbol{z}_t'\boldsymbol{\gamma} + e_t$, where $\boldsymbol{x}_t$ is $k_1 \times 1$ and $\boldsymbol{z}_t$ is $k_2 \times 1$. Suppose that

$$\mathbb{E}(y_t|\mathcal{Y}^{t-1}, \mathcal{W}^t) = \boldsymbol{x}_t'\boldsymbol{\beta}_o.$$

Assuming suitable weak laws for $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$, what are the probability limits of the OLS estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$?

6.4 Consider a linear specification with $\boldsymbol{x}_t = (1\ d_t)'$, where d_t is a one-time dummy: $d_t = 1$ if $t = t^*$, a particular date, and $d_t = 0$ otherwise. What is

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t')?$$

How can you prove (or disprove) OLS consistency?

6.5 State the conditions under which the OLS estimator of seemingly unrelated regressions is consistent and asymptotically normally distributed.

6.6 For Example 6.12, suppose that ϵ_t are serially correlated with unequal variances. Given the asymptotic normality result for the OLS estimators $\hat{a}_T$ and $\hat{b}_T$.

6.7 Given $y_t = \boldsymbol{x}_t'\boldsymbol{\beta}_o$, if $\{\epsilon_t\}$ is a martingale difference sequence with respect to $\{\mathcal{Y}^{t-1}, \mathcal{W}^t\}$, prove that $\mathbb{E}(\epsilon_t) = 0$ and $\mathbb{E}(\epsilon_t\epsilon_\tau) = 0$ for all $t \neq \tau$. Is $\{\epsilon_t\}$ a white noise? Why or why not?

6.8 Given the conditions of Theorem 6.3, let $\epsilon_t = y_t - \boldsymbol{x}_t'\boldsymbol{\beta}_o$ such that $\mathbb{E}(\epsilon_t^2|\boldsymbol{x}_t) = \sigma_o^2$. Prove that the standard OLS variance estimator $\hat{\sigma}_T^2$ is weakly consistent for σ_o^2.

6.9 Prove that under the null hypothesis, $\mathcal{W}_T - \mathcal{LM}_T \xrightarrow{\mathbb{P}} 0$. Also show that when $\boldsymbol{V}_o$ is known, $\mathcal{W}_T = \mathcal{LM}_T$.

References

Andrews, Donald W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation, *Econometrica*, **59**, 817–858.

Davidson, James (1994). *Stochastic Limit Theory*, New York, NY: Oxford University Press.

Gallant, A. Ronald (1987). *Nonlinear Statistical Models*, New York, NY: Wiley.

Eicker, (1967).

Godfrey, L. G. (1988). *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches*, New York, NY: Cambridge University Press.

Newey, Whitney K. and Kenneth West (1987). A simple positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica*, **55**, 703–708.

White, Halbert (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica*, **48**, 817–838.

White, Halbert (1984). *Asymptotic Theory for Econometricians*, Orlando, FL: Academic Press.

Chapter 7

Nonlinear Least Squares Theory

For real world data, it is hard to believe that linear specifications are "universal" in characterizing all economic relationships. A straightforward extension of linear specifications is to consider specifications that are nonlinear in parameters. For example, the function $\alpha+\beta x^\gamma$ offers more flexibility than the simple linear function $\alpha+\beta x$. Although such an extension is quite natural, it also creates various difficulties. First, deciding an appropriate nonlinear function is typically difficult. Second, it is usually cumbersome to estimate nonlinear specifications and analyze the properties of the resulting estimators. Last, but not the least, estimation results of nonlinear specification may not be easily interpreted.

Despite these difficulties, more and more empirical evidences show that many economic relationships are in fact nonlinear. Examples include nonlinear production functions, regime switching in output series, and time series models that can capture asymmetric dynamic patterns. In this chapter, we concentrate on the estimation of and hypothesis testing for nonlinear specifications. For more discussion of nonlinear regressions we refer to Gallant (1987), Gallant and White (1988), Davidson and MacKinnon (1993) and Bierens (1994).

7.1 Nonlinear Specifications

We consider the nonlinear specification

$$y = f(\boldsymbol{x};\boldsymbol{\beta}) + e(\boldsymbol{\beta}), \tag{7.1}$$

where f is a given function with $\boldsymbol{x}$ an $\ell \times 1$ vector of explanatory variables and $\boldsymbol{\beta}$ a $k\times 1$ vector of parameters, and $e(\boldsymbol{\beta})$ denotes the error of the specification. Note that for

a nonlinear specification, the number of explanatory variables ℓ need not be the same as the number of parameters k. This formulation includes the linear specification as a special case with $f(\boldsymbol{x};\boldsymbol{\beta}) = \boldsymbol{x}'\boldsymbol{\beta}$ and $\ell = k$. Clearly, nonlinear functions that can be expressed in a linear form should be treated as linear specifications. For example, a specification involving a structural change is nonlinear in parameters:

$$y_t = \begin{cases} \alpha + \beta x_t + e_t, & t \le t^*, \\ (\alpha + \delta) + \beta x_t + e_t, & t > t^*, \end{cases}$$

but it is equivalent to the linear specification:

$$y_t = \alpha + \delta D_t + \beta x_t + e_t,$$

where $D_t = 0$ if $t \le t^*$ and $D_t = 1$ if $t > t^*$. Our discussion in this chapter focuses on the specifications that cannot be expressed as linear functions.

There are numerous nonlinear specifications considered in empirical applications. A flexible nonlinear specification is

$$y_t = \alpha + \beta \frac{x_t^\gamma - 1}{\gamma} + e_t,$$

where $(x_t^\gamma - 1)/\gamma$ is the so-called Box-Cox transform of x_t, which yields different functions, depending on the value γ. For example, the Box-Cox transform yields $x_t - 1$ when $\gamma = 1$, $1 - 1/x_t$ when $\gamma = -1$, and a value close to $\ln x_t$ when γ approaches zero. This function is thus more flexible than, e.g., the linear specification $\alpha + \beta x$ and nonlinear specification $\alpha + \beta x^\gamma$. Note that the Box-Cox transformation is often applied to positively valued variables.

In the study of firm behavior, the celebrated CES (constant elasticity of substitution) production function suggests characterizing the output y by the following nonlinear function:

$$y = \alpha\left[\delta L^{-\gamma} + (1-\delta)K^{-\gamma}\right]^{-\lambda/\gamma},$$

where L denotes labor, K denotes capital, α, γ, δ and λ are parameters such that $\alpha > 0$, $0 < \delta < 1$ and $\gamma \ge -1$. The elasticity of substitution for a CES production function is

$$s = \frac{\mathrm{d}\ln(K/L)}{\mathrm{d}\ln(\mathrm{MP}_L/\mathrm{MP}_K)} = \frac{1}{(1+\gamma)} \ge 0,$$

where MP denotes marginal product. This function includes the linear, Cobb-Douglas, Leontief production functions as special cases. To estimate the CES production function, the following nonlinear specification is usually considered:

$$\ln y = \ln\alpha - \frac{\lambda}{\gamma}\ln\left[\delta L^{-\gamma} + (1-\delta)K^{-\gamma}\right] + e;$$

for a different estimation strategy, see Exercise 7.3. On the other hand, the translog (transcendental logarithmic) production function is nonlinear in variables but linear in parameters:

$$\ln y = \beta_1 + \beta_2 \ln L + \beta_3 \ln K + \beta_4 (\ln L)(\ln K) + \beta_5 (\ln L)^2 + \beta_6 (\ln K)^2,$$

and hence can be estimated by the OLS method.

In the time series context, a nonlinear AR(p) specification is

$$y_t = f(y_{t-1}, \ldots, y_{t-p}) + e_t.$$

For example, the *exponential autoregressive* (EXPAR) specification takes the following form:

$$y_t = \sum_{j=1}^{p} \left[\alpha_j + \beta_j \exp\left(-\gamma y_{t-1}^2\right)\right] y_{t-j} + e_t,$$

where in some cases one may replace y_{t-1}^2 in the exponential function with y_{t-j}^2 for $j = 1, \ldots, p$. This specification was designed to describe physical vibration whose amplitude depends on the magnitude of y_{t-1}.

As another example, consider the *self-exciting threshold autoregressive* (SETAR) specification:

$$y_t = \begin{cases} a_0 + a_1 y_{t-1} + \cdots + a_p y_{t-p} + e_t, & \text{if } y_{t-d} \in (-\infty, c], \\ b_0 + b_1 y_{t-1} + \cdots + b_p y_{t-p} + e_t, & \text{if } y_{t-d} \in (c, \infty), \end{cases}$$

where d is known as the "delay parameter" which is an integer between 1 and p, and c is the "threshold parameter." Note that the SETAR model is different from the structural change model in that the parameters switch from one regime to another depending on whether a past realization y_{t-d} exceeds the threshold value c. This specification can be easily extended to allow for r threshold parameters, so that the specification switches among $r + 1$ different dynamic structures.

The SETAR specification above can be written as

$$y_t = a_0 + \sum_{j=1}^{p} a_j y_{t-j} + \left(\Delta_0 + \sum_{j=1}^{p} \Delta_j y_{t-j}\right) \mathbf{1}_{\{y_{t-d} > c\}} + e_t,$$

where $a_j + \Delta_j = b_j$, and $\mathbf{1}$ denotes the indicator function. To avoid abrupt changes of parameters, one may replace the indicator function with a "smooth" function h so as

to allow for smoother transitions of structures. It is typical to choose the function h as a distribution function, e.g.,

$$h(y_{t-d}; c, \delta) = \frac{1}{1 + \exp[-(y_{t-d} - c)/\delta]},$$

where c is still the threshold value and δ is a scale parameter. This leads to the following *smooth threshold autoregressive* (STAR) specification:

$$y_t = a_0 + \sum_{j=1}^{p} a_j y_{t-j} + \left(\Delta_0 + \sum_{j=1}^{p} \Delta_j y_{t-j}\right) h(y_{t-d}; c, \delta) + e_t.$$

Clearly, this specification behaves similarly to a SETAR specification when $|(y_{t-d} - c)/\delta|$ is very large. For more nonlinear time series models and their motivations we refer to Tong (1990).

Another well known nonlinear specification is the so-called *artificial neural network* which has been widely used in cognitive science, engineering, biology and linguistics. A 3-layer neural network can be expressed as

$$f(x_1 \ldots, x_p; \boldsymbol{\beta}) = g\left(\alpha_0 + \sum_{i=1}^{q} \alpha_i \, h\Big(\gamma_{i0} + \sum_{j=1}^{p} \gamma_{ij} x_j\Big)\right),$$

where $\boldsymbol{\beta}$ is the parameter vector containing all α and γ, g and h are some pre-specified functions. In the jargon of the neural network literature, this specification contains p "inputs units" in the input layer (each corresponding to an explanatory variable x_j), q "hidden units" in the hidden (middle) layer with the i^{th} hidden-unit activation $h_i = h(\gamma_{i0} + \sum_{j=1}^{p} \gamma_{ij} x_j)$, and one "output unit" in the output layer with the activation $o = g(\beta_0 + \sum_{i=1}^{q} \beta_i h_i)$. The functions h and g are known as "activation functions," the parameters in these functions are "connection weights." That is, the input values simultaneously activate q hidden units, and these hidden-unit activations in turn determine the output value. The output value is supposed to capture the behavior of the "target" (dependent) variable y. In the context of nonlinear regression, we can write

$$y = g\left(\alpha_0 + \sum_{i=1}^{q} \alpha_i \, h\Big(\gamma_{i0} + \sum_{j=1}^{p} \gamma_{ij} x_j\Big)\right) + e,$$

For a multivariate target $\boldsymbol{y}$, networks with multiple outputs can be constructed similarly with g being a vector-valued function.

In practice, it is typical to choose h as a "sigmoid" (S-shaped) function bounded within a certain range. For example, two leading choices of h are the logistic function

$h(x) = 1/(1 + e^{-x})$ which is bounded between 0 and 1 and the hyperbolic tangent function

$$h(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}},$$

which is bounded between -1 and 1. The function g may be the identity function or the same as h. Although the class of neural networks is highly nonlinear in parameters, it possesses two appealing properties. First, a neural network is capable of approximating any Borel-measurable function to any degree of accuracy, provided that the number of hidden units q is sufficiently large. Second, to achieve a given degree of approximation accuracy, neural networks are relatively more parsimonious than, e.g., the polynomial and trignometric expansions. For more details of artificial neural networks and their relationships to econometrics we refer to Kuan and White (1994).

7.2 The Method of Nonlinear Least Squares

Formally, we consider the nonlinear specification (7.1):

$$y = f(\boldsymbol{x}; \boldsymbol{\beta}) + e(\boldsymbol{\beta}),$$

where $f\colon \mathbb{R}^\ell \times \Theta_1 \mapsto \mathbb{R}$, Θ_1 denotes the parameter space, a subspace of $\mathbb{R}^k$, and $e(\boldsymbol{\beta})$ is the specification error. Given T observations of y and $\boldsymbol{x}$, let

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}, \qquad \boldsymbol{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T; \boldsymbol{\beta}) = \begin{bmatrix} f(\boldsymbol{x}_1; \boldsymbol{\beta}) \\ f(\boldsymbol{x}_2; \boldsymbol{\beta}) \\ \vdots \\ f(\boldsymbol{x}_T; \boldsymbol{\beta}) \end{bmatrix}.$$

The nonlinear specification (7.1) now can be expressed as

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T; \boldsymbol{\beta}) + \boldsymbol{e}(\boldsymbol{\beta}),$$

where $\boldsymbol{e}(\boldsymbol{\beta})$ is the vector of errors.

7.2.1 Nonlinear Least Squares Estimator

Our objective is to find a k-dimensional surface that "best" fits the data $(y_t, \boldsymbol{x}_t)$, $t = 1, \ldots, T$. Analogous to the OLS method, the method of *nonlinear least squares* (NLS)

suggests to minimize the following NLS criterion function with respect to $\boldsymbol{\beta}$:

$$\begin{aligned} Q_T(\boldsymbol{\beta}) &= \frac{1}{T}[\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T; \boldsymbol{\beta})]'[\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T; \boldsymbol{\beta})] \\ &= \frac{1}{T}\sum_{t=1}^{T}[y_t - f(\boldsymbol{x}_t; \boldsymbol{\beta})]^2. \end{aligned} \tag{7.2}$$

Note that Q_T is also a function of the data y_t and x_t; we omit the arguments y_t and $\boldsymbol{x}_t$ just for convenience.

The first order condition of the NLS minimization problem is a system of k nonlinear equations with k unknowns:

$$\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}) = -\frac{2}{T}\nabla_{\boldsymbol{\beta}} \boldsymbol{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T; \boldsymbol{\beta})\,[\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T; \boldsymbol{\beta})] \overset{\text{set}}{=} \boldsymbol{0},$$

where

$$\nabla_{\boldsymbol{\beta}} \boldsymbol{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T; \boldsymbol{\beta}) = \left[\begin{array}{cccc} \nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_1; \boldsymbol{\beta}) & \nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_2; \boldsymbol{\beta}) & \ldots & \nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_T; \boldsymbol{\beta}) \end{array}\right],$$

is a $k \times T$ matrix. A solution to this minimization problem is some $\bar{\boldsymbol{\beta}} \in \Theta_1$ that solves the first order condition: $\nabla_{\boldsymbol{\beta}} Q_T(\bar{\boldsymbol{\beta}}) = 0$, and satisfies the second order condition: $\nabla^2_{\boldsymbol{\beta}} Q_T(\bar{\boldsymbol{\beta}})$ is positive definite. We thus impose the following identification requirement; cf. [ID-1] for linear specifications.

[ID-2] $f(\boldsymbol{x}; \cdot)$ is twice continuously differentiable in the second argument on Θ_1, such that for given data $(y_t, \boldsymbol{x}_t)$, $t = 1, \ldots, T$, $\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta})$ is positive definite at some interior point of Θ_1.

While [ID-2] ensures that a minimum of $Q_T(\boldsymbol{\beta})$ can be found, it does not guarantee the uniqueness of this solution. For a a given data set, there may exist multiple solutions to the NLS minimization problem such that each solution is a local minimum of $Q_T(\boldsymbol{\beta})$. This result is stated below; cf. Theorem 3.1.

Theorem 7.1 *Given the specification* (7.1)*, suppose that* [ID-2] *holds. Then, there exists a solution that minimizes the NLS criterion function* (7.2)*.*

Writing $\boldsymbol{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T; \boldsymbol{\beta})$ as $\boldsymbol{f}(\boldsymbol{\beta})$, we have

$$\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}) = -\frac{2}{T}\nabla^2_{\boldsymbol{\beta}} \boldsymbol{f}(\boldsymbol{\beta})\,[\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\beta})] + \frac{2}{T}[\nabla_{\boldsymbol{\beta}} \boldsymbol{f}(\boldsymbol{\beta})][\nabla_{\boldsymbol{\beta}} \boldsymbol{f}(\boldsymbol{\beta})]'.$$

For linear regressions, $\boldsymbol{f}(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$ so that $\nabla_{\boldsymbol{\beta}} \boldsymbol{f}(\boldsymbol{\beta}) = \boldsymbol{X}'$ and $\nabla^2_{\boldsymbol{\beta}} \boldsymbol{f}(\boldsymbol{\beta}) = \boldsymbol{0}$. It follows that $\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}) = 2(\boldsymbol{X}'\boldsymbol{X})/T$, which is positive definite if, and only if, $\boldsymbol{X}$ has full

column rank. This shows that [ID-2] is, in effect, analogous to [ID-1] for the OLS method. Comparing to the OLS method, the NLS minimization problem may not have a closed-form solution because the first order condition is a system of nonlinear functions in general; see also Exercise 7.1.

The minimizer of $Q_T(\boldsymbol{\beta})$ is known as the NLS estimator and will be denoted as $\hat{\boldsymbol{\beta}}_T$. Let $\hat{\boldsymbol{y}}$ denote the vector of NLS fitted values with the t^{th} element $\hat{y}_t = f(\boldsymbol{x}_t, \hat{\boldsymbol{\beta}}_T)$, and $\hat{\boldsymbol{e}}$ denote the vector of NLS residuals $\boldsymbol{y} - \hat{\boldsymbol{y}}$ with the t^{th} element $\hat{e}_t = y_t - \hat{y}_t$. Denote the transpose of $\nabla_{\boldsymbol{\beta}} \boldsymbol{f}(\boldsymbol{\beta})$ as $\boldsymbol{\Xi}(\boldsymbol{\beta})$. Then by the first order condition,

$$\boldsymbol{\Xi}(\hat{\boldsymbol{\beta}}_T)'\hat{\boldsymbol{e}} = [\nabla_{\boldsymbol{\theta}} \boldsymbol{f}(\hat{\boldsymbol{\beta}}_T)]\hat{\boldsymbol{e}} = \boldsymbol{0}.$$

That is, the residual vector is orthogonal to every column vector of $\boldsymbol{\Xi}(\hat{\boldsymbol{\beta}}_T)$. Geometrically, $\boldsymbol{f}(\boldsymbol{\beta})$ defines a surface on Θ_1, and for any $\boldsymbol{\beta}$ in Θ_1, $\boldsymbol{\Xi}(\boldsymbol{\beta})$ is a k-dimensional linear subspace tangent at the point $\boldsymbol{f}(\boldsymbol{\beta})$. Thus, $\boldsymbol{y}$ is orthogonally projected onto this surface at $\boldsymbol{f}(\hat{\boldsymbol{\beta}}_T)$ so that the residual vector is orthogonal to the tangent space at that point. In contrast with linear regressions, there may be more than one orthogonal projections and hence multiple solutions to the NLS minimization problem. There is also no guarantee that the sum of NLS residuals is zero; see Exercise 7.2.

Remark: The marginal response to the change of the i^{th} regressor is $\partial f(\boldsymbol{x}_t; \boldsymbol{\beta})/\partial x_{ti}$. Thus, one should be careful in interpreting the estimation results because a parameter in a nonlinear specification is not necessarily the marginal response to the change of a regressor.

7.2.2 Nonlinear Optimization Algorithms

When a solution to the first order condition of the NLS minimization problem cannot be obtained analytically, the NLS estimates must be computed using numerical methods. To optimizing a nonlinear function, an *iterative algorithm* starts from some initial value of the argument in that function and then repeatedly calculates next available value according to a particular rule until an optimum is reached approximately. It should be noted that when there are multiple optima, an iterative algorithm may not be able to locate the global optimum. In fact, it is more common that an algorithm gets stuck at a local optimum, except in some special cases, e.g., when optimizing a globally concave (convex) function. In the literature, several new methods, such as the *simulated annealing algorithm*, have been proposed to find the global solution. These methods have not yet been standard because they are typically difficult to implement and computationally very intensive. We will therefore confine ourselves to those commonly used "local"

methods.

To minimize $Q_T(\boldsymbol{\beta})$, a generic algorithm can be expressed as

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} + s^{(i)}\boldsymbol{d}^{(i)},$$

so that the $(i+1)^{\text{th}}$ iterated value $\boldsymbol{\beta}^{(i+1)}$ is obtained from $\boldsymbol{\beta}^{(i)}$, the value from the previous iteration, by adjusting the amount $s^{(i)}\boldsymbol{d}^{(i)}$, where $\boldsymbol{d}^{(i)}$ characterizes the direction of change in the parameter space and $s^{(i)}$ controls the amount of change. Different algorithms are resulted from different choices of s and $\boldsymbol{d}$. As maximizing Q_T is equivalent to minimizing $-Q_T$, the methods discussed here are readily modified to the algorithms for maximization problems.

Consider the first-order Taylor expansion of $Q(\boldsymbol{\beta})$ about $\boldsymbol{\beta}^\dagger$:

$$Q_T(\boldsymbol{\beta}) \approx Q_T(\boldsymbol{\beta}^\dagger) + [\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^\dagger)]'(\boldsymbol{\beta} - \boldsymbol{\beta}^\dagger).$$

Replacing $\boldsymbol{\beta}$ with $\boldsymbol{\beta}^{(i+1)}$ and $\boldsymbol{\beta}^\dagger$ with $\boldsymbol{\beta}^{(i)}$ we have

$$Q_T\big(\boldsymbol{\beta}^{(i+1)}\big) \approx Q_T\big(\boldsymbol{\beta}^{(i)}\big) + \big[\nabla_{\boldsymbol{\beta}} Q_T\big(\boldsymbol{\beta}^{(i)}\big)\big]' s^{(i)}\boldsymbol{d}^{(i)}.$$

Note that this approximation is valid when $\boldsymbol{\beta}^{(i+1)}$ is in the neighborhood of $\boldsymbol{\beta}^{(i)}$. Let $\boldsymbol{g}(\boldsymbol{\beta})$ denote the gradient vector of Q_T: $\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta})$, and $\boldsymbol{g}^{(i)}$ denote $\boldsymbol{g}(\boldsymbol{\beta})$ evaluated at $\boldsymbol{\beta}^{(i)}$. If $\boldsymbol{d}^{(i)} = -\boldsymbol{g}^{(i)}$,

$$Q_T\big(\boldsymbol{\beta}^{(i+1)}\big) \approx Q_T\big(\boldsymbol{\beta}^{(i)}\big) - s^{(i)}\big[\boldsymbol{g}^{(i)\prime}\boldsymbol{g}^{(i)}\big].$$

As $\boldsymbol{g}^{(i)\prime)}\boldsymbol{g}^{(i)}$ is non-negative, we can find a positive and small enough s such that Q_T is decreasing. Clearly, when $\boldsymbol{\beta}^{(i)}$ is already a minimum of Q_T, $\boldsymbol{g}^{(i)}$ is zero so that no further adjustment is possible. This suggests the following algorithm:

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} - s^{(i)}\boldsymbol{g}^{(i)}.$$

Choosing $\boldsymbol{d}^{(i)} = \boldsymbol{g}^{(i)}$ leads to:

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} + s^{(i)}\boldsymbol{g}^{(i)},$$

which can be used to search for a maximum of Q_T.

Given the search direction, one may want to choose $s^{(i)}$ such that the next value of the objective function $Q_T\big(\boldsymbol{\beta}^{(i+1)}\big)$ is a minimum. This suggests that the first order condition below should hold:

$$\frac{\partial Q_T\big(\boldsymbol{\beta}^{(i+1)}\big)}{\partial s^{(i)}} = \nabla_{\boldsymbol{\beta}} Q_T\big(\boldsymbol{\beta}^{(i+1)}\big)\,\frac{\partial \boldsymbol{\beta}^{(i+1)}}{\partial s^{(i)}} = -\boldsymbol{g}^{(i+1)\prime}\boldsymbol{g}^{(i)} = 0.$$

Let $\boldsymbol{H}^{(i)}$ denote the Hessian matrix of Q_T evaluated at $\boldsymbol{\beta}^{(i)}$:

$$\boldsymbol{H}^{(i)} = \nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(i)}} = \nabla_{\boldsymbol{\beta}}\, g(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(i)}}.$$

Then by Taylor's expansion of g, we have

$$\boldsymbol{g}^{(i+1)} \approx \boldsymbol{g}^{(i)} + \boldsymbol{H}^{(i)}\big(\boldsymbol{\beta}^{(i+1)} - \boldsymbol{\beta}^{(i)}\big) = \boldsymbol{g}^{(i)} - \boldsymbol{H}^{(i)} s^{(i)} \boldsymbol{g}^{(i)}.$$

It follows that

$$0 = \boldsymbol{g}^{(i+1)\prime}\boldsymbol{g}^{(i)} \approx \boldsymbol{g}^{(i)\prime}\boldsymbol{g}^{(i)} - s^{(i)}\boldsymbol{g}^{(i)\prime}\boldsymbol{H}^{(i)}\boldsymbol{g}^{(i)},$$

or equivalently,

$$s^{(i)} = \frac{\boldsymbol{g}^{(i)\prime}\boldsymbol{g}^{(i)}}{\boldsymbol{g}^{(i)\prime}\boldsymbol{H}^{(i)}\boldsymbol{g}^{(i)}}.$$

The step length $s^{(i)}$ is non-negative whenever $\boldsymbol{H}^{(i)}$ is positive definite. The algorithm derived above now reads

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} - \frac{\boldsymbol{g}^{(i)\prime}\boldsymbol{g}^{(i)}}{\boldsymbol{g}^{(i)\prime}\boldsymbol{H}^{(i)}\boldsymbol{g}^{(i)}}\,\boldsymbol{g}^{(i)},$$

which is known as the *steepest descent algorithm*. If $H^{(i)}$ is not positive definite, $s^{(i)}$ may be non-negative so that this algorithm may point to a wrong direction.

As the steepest descent algorithm adjusts parameters along the opposite of the gradient direction, it may run into difficulty when, e.g., the nonlinear function being optimized is flat around the optimum. The algorithm may iterate back and forth without much progress in approaching an optimum. An alternative is to consider the second-order Taylor expansion of $Q(\boldsymbol{\beta})$ around some $\boldsymbol{\beta}^\dagger$:

$$Q_T(\boldsymbol{\beta}) \approx Q_T(\boldsymbol{\beta}^\dagger) + \boldsymbol{g}^{\dagger\prime}(\boldsymbol{\beta} - \boldsymbol{\beta}^\dagger) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^\dagger)'\boldsymbol{H}^\dagger(\boldsymbol{\beta} - \boldsymbol{\beta}^\dagger),$$

where $\boldsymbol{g}^\dagger$ and $\boldsymbol{H}^\dagger$ are $\boldsymbol{g}$ and $\boldsymbol{H}$ evaluated at $\boldsymbol{\beta}^\dagger$, respectively. From this expansion, the first order condition of $Q_T(\boldsymbol{\beta})$ may be expressed as

$$\boldsymbol{g}^\dagger + \boldsymbol{H}^\dagger(\boldsymbol{\beta} - \boldsymbol{\beta}^\dagger) \approx \boldsymbol{0},$$

so that $\boldsymbol{\beta} \approx \boldsymbol{\beta}^\dagger - (\boldsymbol{H}^\dagger)^{-1}\boldsymbol{g}^\dagger$. This suggests the following algorithm:

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} - \big(\boldsymbol{H}^{(i)}\big)^{-1}\boldsymbol{g}^{(i)},$$

where the step length is 1, and the direction vector is $-\big(\boldsymbol{H}^{(i)}\big)^{-1}\boldsymbol{g}^{(i)}$. This is also known as the *Newton-Raphson algorithm*. This algorithm is more difficult to implement because it involves matrix inversion at each iteration step.

From Taylor's expansion we can also see that

$$Q_T\big(\boldsymbol{\beta}^{(i+1)}\big) - Q_T\big(\boldsymbol{\beta}^{(i)}\big) \approx -\frac{1}{2}\,\boldsymbol{g}^{(i)\prime}\big(\boldsymbol{H}^{(i)}\big)^{-1}\boldsymbol{g}^{(i)},$$

where the right-hand side is negative provided that $\boldsymbol{H}^{(i)}$ is positive definite. When this approximation is good, the Newton-Raphson algorithm usually (but not always) results in a decrease in the value of Q_T. This algorithm may point to a wrong direction if $\boldsymbol{H}^{(i)}$ is not positive definite; this happens when, e.g., Q is concave at $\boldsymbol{\beta}^i$. When Q_T is (locally) quadratic with the local minimum $\boldsymbol{\beta}^*$, the second-order expansion about $\boldsymbol{\beta}^*$ is exact, and hence

$$\boldsymbol{\beta} = \boldsymbol{\beta}^* - \boldsymbol{H}(\boldsymbol{\beta}^*)^{-1}\boldsymbol{g}(\boldsymbol{\beta}^*).$$

In this case, the Newton-Raphson algorithm can reach the minimum in a single step. Alternatively, we may also add a step length to the Newton-Raphson algorithm:

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} - s^{(i)}\big(\boldsymbol{H}^{(i)}\big)^{-1}\boldsymbol{g}^{(i)},$$

where $s^{(i)}$ may be found by minimizing $Q\big(\boldsymbol{\beta}^{(i+1)}\big)$. In practice, it is more typical to choose $s^{(i)}$ such that $Q\big(\boldsymbol{\beta}^{(i)}\big)$ is decreasing at each iteration.

A algorithm that avoids computing the second-order derivatives is the so-called *Gauss-Newton algorithm*. When $Q_T(\boldsymbol{\beta})$ is the NLS criterion function,

$$\boldsymbol{H}(\boldsymbol{\beta}) = -\frac{2}{T}\nabla^2_{\boldsymbol{\beta}}\boldsymbol{f}(\boldsymbol{\beta})[\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\beta})] + \frac{2}{T}\boldsymbol{\Xi}(\boldsymbol{\beta})'\boldsymbol{\Xi}(\boldsymbol{\beta}),$$

where $\boldsymbol{\Xi}(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}}\boldsymbol{f}(\boldsymbol{\beta})$. It is therefore convenient to ignore the first term on the right-hand side and approximate $\boldsymbol{H}(\boldsymbol{\beta})$ by $2\boldsymbol{\Xi}(\boldsymbol{\beta})'\boldsymbol{\Xi}(\boldsymbol{\beta})/T$. There are some advantages of this approximation. First, only the first-order derivatives need to be computed. Second, this approximation is guaranteed to be positive definite under [ID-2]. The resulting algorithm is

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} + \big[\boldsymbol{\Xi}\big(\boldsymbol{\beta}^{(i)}\big)'\boldsymbol{\Xi}\big(\boldsymbol{\beta}^{(i)}\big)\big]^{-1}\boldsymbol{\Xi}\big(\boldsymbol{\beta}^{(i)}\big)\big[\boldsymbol{y} - \boldsymbol{f}\big(\boldsymbol{\beta}^{(i)}\big)\big].$$

Observe that the adjustment term can be obtained as the OLS estimator of regressing $\boldsymbol{y} - \boldsymbol{f}\big(\boldsymbol{\beta}^{(i)}\big)$ on $\boldsymbol{\Xi}\big(\boldsymbol{\beta}^{(i)}\big)$; this regression is thus known as the *Gauss-Newton regression*. The iterated $\boldsymbol{\beta}$ values can be easily computed by performing the Gauss-Newton regression repeatedly. The performance of this algorithm may be quite different from the

Newton-Raphson algorithm because it utilizes only an approximation to the Hessian matrix.

To maintain a correct search direction of the steepest descent and Newton-Raphson algorithms, it is important to ensure that $\boldsymbol{H}^{(i)}$ is positive definite at each iteration. A simple approach is to correct $\boldsymbol{H}^{(i)}$, if necessary, by adding an appropriate matrix to it. A popular correction is

$$\boldsymbol{H}_c^{(i)} = \boldsymbol{H}^{(i)} + c^{(i)}\boldsymbol{I},$$

where $c^{(i)}$ is a positive number chosen to "force" $\boldsymbol{H}_c^{(i)}$ to be a positive definite matrix. Let $\tilde{\boldsymbol{H}} = \boldsymbol{H}^{-1}$. One may also compute

$$\tilde{\boldsymbol{H}}_c^{(i)} = \tilde{\boldsymbol{H}}^{(i)} + c\boldsymbol{I},$$

because it is the inverse of $\boldsymbol{H}^{(i)}$ that matters in the algorithm. Such a correction is used in, for example, the so-called *Marquardt-Levenberg algorithm*.

The *quasi-Newton method*, on the other hand, corrects $\tilde{\boldsymbol{H}}^{(i)}$ iteratively by adding a symmetric, correction matrix $C^{(i)}$:

$$\tilde{\boldsymbol{H}}^{(i+1)} = \tilde{\boldsymbol{H}}^{(i)} + \boldsymbol{C}^{(i)},$$

with the initial value $\tilde{\boldsymbol{H}}^{(0)} = \boldsymbol{I}$. This method includes the Davidon-Fletcher-Powell (DFP) algorithm and the Broydon-Fletcher-Goldfarb-Shanno (BFGS) algorithm, where the latter is the algorithm used in the GAUSS program. In the DFP algorithm,

$$\boldsymbol{C}^{(i)} = \frac{\boldsymbol{\delta}^{(i)}\boldsymbol{\delta}^{(i)\prime}}{\boldsymbol{\delta}^{(i)\prime}\boldsymbol{\gamma}^{(i)}} + \frac{\tilde{\boldsymbol{H}}^{(i)}\boldsymbol{\gamma}^{(i)}\boldsymbol{\gamma}^{(i)\prime}\tilde{\boldsymbol{H}}^{(i)}}{\boldsymbol{\gamma}^{(i)\prime}\tilde{\boldsymbol{H}}^{(i)}\boldsymbol{\gamma}^{(i)}},$$

where $\boldsymbol{\delta}^{(i)} = \boldsymbol{\beta}^{(i+1)} - \boldsymbol{\beta}^{(i)}$ and $\boldsymbol{\gamma}^{(i)} = \boldsymbol{g}^{(i+1)} - \boldsymbol{g}^{(i)}$. The BFGS algorithm contains an additional term in the correction matrix.

To implement an iterative algorithm, one must choose a vector of initial values to start the algorithm and a stopping rule to terminate the iteration procedure. Initial values are usually specified by the researcher or by random number generation; prior information, if available, should also be taken into account. For example, if the parameter is a probability, the algorithm may be initialized by, say, 0.5 or by a number randomly generated from the uniform distribution on $[0, 1]$. Without prior information, it is also typical to generate initial values from a normal distribution. In practice, one would generate many sets of initial values and then choose the one that leads to a better result

(for example, a better fit of data). Of course, this search process is computationally demanding.

When an algorithm results in no further improvement, a stopping rule must be invoked to terminate the iterations. Typically, an algorithm stops when one of the following convergence criteria is met: for a pre-determined, small positive number c,

1. $\left\|\boldsymbol{\beta}^{(i+1)} - \boldsymbol{\beta}^{(i)}\right\| < c$, where $\|\cdot\|$ denotes the Euclidean norm,
2. $\left\|\boldsymbol{g}\left(\boldsymbol{\beta}^{(i)}\right)\right\| < c$, or
3. $\left|Q_T\left(\boldsymbol{\beta}^{(i+1)}\right) - Q_T\left(\boldsymbol{\beta}^{(i)}\right)\right| < c$.

For the Gauss-Newton algorithm, one may stop the algorithm when TR^2 is "close" to zero, where R^2 is the coefficient of determination of the Gauss-Newton regression. As the residual vector must be orthogonal to the tangent space at the optimum, this stopping rule amounts to checking whether the first order condition is satisfied approximately. In some cases, an algorithm may never meet its pre-set convergence criterion and hence keeps on iterating. To circumvent this difficulty, an optimization program usually sets a maximum number for iterations so that the program terminates automatically once the number of iterations reaches this upper bound.

7.3 Asymptotic Properties of the NLS Estimators

7.3.1 Consistency

As the NLS estimator does not have an analytic form in general, a different approach is thus needed to establish NLS consistency. Intuitively, when the NLS objective function $Q_T(\boldsymbol{\beta})$ is close to $\mathbb{E}[Q_T(\boldsymbol{\beta})]$ for all $\boldsymbol{\beta}$, it is reasonable to expect that the minimizer of $Q_T(\boldsymbol{\beta})$, i.e., the NLS estimator $\hat{\boldsymbol{\beta}}_T$, is also close to a minimum of $\mathbb{E}[Q_T(\boldsymbol{\beta})]$. Given that Q_T is nonlinear in $\boldsymbol{\beta}$, a ULLN must be invoked to justify the closeness between $Q_T(\boldsymbol{\beta})$ and $\mathbb{E}[Q_T(\boldsymbol{\beta})]$, as discussed in Section 5.6.

To illustrate how consistency can be obtained, we consider a special case. Suppose that $\mathbb{E}[Q_T(\boldsymbol{\beta})]$ is a continuous function on the compact parameter space Θ_1 such that $\boldsymbol{\beta}_o$ is its unique, global minimum. The NLS estimator $\hat{\boldsymbol{\beta}}_T$ is such that

$$Q_T(\hat{\boldsymbol{\beta}}_T) = \inf_{\Theta_1} Q_T(\boldsymbol{\beta}).$$

Suppose also that Q_T has a SULLN effect, i.e., there is a set $\Omega_0 \subseteq \Omega$ such that $\mathbb{P}(\Omega_0) = 1$ and

$$\sup_{\boldsymbol{\beta}\in\Theta_1} \left|Q_T(\boldsymbol{\beta}) - \mathbb{E}[Q_T(\boldsymbol{\beta})]\right| \to 0,$$

for all $\omega \in \Omega_0$. Set

$$\epsilon = \inf_{\boldsymbol{\beta}\in B^c\cap\Theta_1} \big(\mathbb{E}[Q_T(\boldsymbol{\beta})] - \mathbb{E}[Q_T(\boldsymbol{\beta}_o)]\big),$$

where B is an open neighborhood of $\boldsymbol{\beta}_o$. Then for $\omega \in \Omega_0$, we can choose T sufficiently large such that

$$\mathbb{E}[Q_T(\hat{\boldsymbol{\beta}}_T)] - Q_T(\hat{\boldsymbol{\beta}}_T) < \frac{\epsilon}{2},$$

and that

$$Q_T(\hat{\boldsymbol{\beta}}_T) - E[Q_T(\boldsymbol{\beta}_o)] \le Q_T(\boldsymbol{\beta}_o) - E[Q_T(\boldsymbol{\beta}_o)] < \frac{\epsilon}{2},$$

because the NLS estimator $\hat{\boldsymbol{\beta}}_T$ minimizes $Q_T(\boldsymbol{\beta})$. It follows that for $\omega \in \Omega_0$,

$$\begin{aligned}
&\mathbb{E}[Q_T(\hat{\boldsymbol{\beta}}_T)] - \mathbb{E}[Q_T(\boldsymbol{\beta}_o)] \\
&\qquad \le \mathbb{E}[Q_T(\hat{\boldsymbol{\beta}}_T)] - Q_T(\hat{\boldsymbol{\beta}}_T) + Q_T(\hat{\boldsymbol{\beta}}_T) - E[Q_T(\boldsymbol{\beta}_o)] \\
&\qquad < \epsilon,
\end{aligned}$$

for all T sufficiently large. This shows that, comparing to all $\boldsymbol{\beta}$ outside the neighborhood B of $\boldsymbol{\beta}_o$, $\hat{\boldsymbol{\beta}}_T$ will eventually render $\mathbb{E}[Q_T(\boldsymbol{\beta})]$ closer to $\mathbb{E}[Q_T(\boldsymbol{\beta}_o)]$ with probability one. Thus, $\hat{\boldsymbol{\beta}}_T$ must be in B for large T. As B is arbitrary, $\hat{\boldsymbol{\beta}}_T$ must converge to $\boldsymbol{\beta}_o$ almost surely. Convergence in probability of $\hat{\boldsymbol{\beta}}_T$ to $\boldsymbol{\beta}_o$ can be established using a similar argument; see e.g., Amemiya (1985) and Exercise 7.4.

The preceding discussion shows what matters for consistency is the effect of a SULLN (WULLN). Recall from Theorem 5.34 that, to ensure a SULLN (WULLN), Q_T should obey a SLLN (WLLN) for each $\boldsymbol{\beta} \in \Theta_1$ and also satisfy a Lipschitz-type continuity condition:

$$|Q_T(\boldsymbol{\beta}) - Q_T(\boldsymbol{\beta}^\dagger)| \le C_T\|\boldsymbol{\beta} - \boldsymbol{\beta}^\dagger\| \quad \text{a.s.},$$

with C_T bounded almost surely (in probability). If the parameter space Θ_1 is compact and convex, we have from the mean-value theorem and the Cauchy-Schwartz inequality that

$$|Q_T(\boldsymbol{\beta}) - Q_T(\boldsymbol{\beta}^\dagger)| \le \|\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^*)\|\,\|\boldsymbol{\beta} - \boldsymbol{\beta}^\dagger\| \quad \text{a.s.},$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^\dagger$ are in Θ_1 and $\boldsymbol{\beta}^*$ is the mean value of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^\dagger$, in the sense that $|\boldsymbol{\beta}^* - \boldsymbol{\beta}_o| < |\boldsymbol{\beta}^\dagger - \boldsymbol{\beta}_o|$. Hence, the Lipschitz-type condition would hold by setting

$$C_T = \sup_{\boldsymbol{\beta}\in\Theta_1} \nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}).$$

Observe that in the NLS context,

$$Q_T(\boldsymbol{\beta}) = \frac{1}{T}\sum_{t=1}^{T}\big(y_t^2 - 2y_t f(\boldsymbol{x}_t;\boldsymbol{\beta}) + f(\boldsymbol{x}_t;\boldsymbol{\beta})^2\big),$$

and

$$\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}) = -\frac{2}{T}\sum_{t=1}^{T}\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta})[y_t - f(\boldsymbol{x}_t;\boldsymbol{\beta})].$$

Hence, $\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta})$ cannot be almost surely bounded in general. (It would be bounded if, for example, y_t are bounded random variables and both f and $\nabla_{\boldsymbol{\beta}} f$ are bounded functions.) On the other hand, it is practically more plausible that $\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta})$ is bounded in probability. It is the case when, for example, $\mathbb{E}\,|\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta})|$ is bounded uniformly in $\boldsymbol{\beta}$. As such, we shall restrict our discussion below to WULLN and weak consistency of $\hat{\boldsymbol{\beta}}_T$.

To proceed we assume that the identification requirement [ID-2] holds with probability one. The discussion above motivates the additional conditions given below.

[C1] $\{(y_t \;\; \boldsymbol{w}_t')'\}$ is a sequence of random vectors, and $\boldsymbol{x}_t$ is vector containing some elements of $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$.

- (i) The sequences $\{y_t^2\}$, $\{y_t f(\boldsymbol{x}_t;\boldsymbol{\beta})\}$ and $\{f(\boldsymbol{x}_t;\boldsymbol{\beta})^2\}$ all obey a WLLN for each $\boldsymbol{\beta}$ in Θ_1, where Θ_1 is compact and convex.
- (ii) y_t, $f(\boldsymbol{x}_t;\boldsymbol{\beta})$ and $\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta})$ all have bounded second moment uniformly in $\boldsymbol{\beta}$.

[C2] There exists a unique parameter vector $\boldsymbol{\beta}_o$ such that $\mathbb{E}(y_t \mid \mathcal{Y}^{t-1}, \mathcal{W}^t) = f(\boldsymbol{x}_t;\boldsymbol{\beta}_o)$.

Condition [C1] is analogous to [B1] so that stochastic regressors are allowed. [C1](i) regulates that each components of $Q_T(\boldsymbol{\beta})$ obey a standard WLLN. [C1](ii) implies

$$\mathbb{E}\,|\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta})| \le \frac{2}{T}\sum_{t=1}^{T}\Big(\|\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta})\|_2\|y_t\|_2 + \|\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta})\|_2\|f(\boldsymbol{x}_t;\boldsymbol{\beta})\|_2\Big) \le \Delta,$$

for some Δ which does not depend on $\boldsymbol{\beta}$. This in turn implies $\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta})$ is bounded in probability (uniformly in $\boldsymbol{\beta}$) by Markov's inequality. Condition [C2] is analogous to [B2] and requires $f(\boldsymbol{x}_t;\boldsymbol{\beta})$ been a correct specification of the conditional mean function. Thus, $\boldsymbol{\beta}_o$ globally minimizes $\mathbb{E}[Q_T(\boldsymbol{\beta})]$ because the conditional mean must minimizes mean-squared errors.

Theorem 7.2 *Given the nonlinear specification (7.1), suppose that* [C1] *and* [C2] *hold. Then,* $\hat{\boldsymbol{\beta}}_T \xrightarrow{\mathbb{P}} \boldsymbol{\beta}_o$.

Theorem 7.2 is not completely satisfactory because it is concerned with the convergence to the global minimum. As noted in Section 7.2.2, an iterative algorithm is not guaranteed to find a global minimum of the NLS objective function. Hence, it is more reasonable to expect that the NLS estimator only converges to some local minimum of $\mathbb{E}[Q_T(\boldsymbol{\beta})]$. A simple proof of such local consistency result is not yet available. We therefore omit the details and assert only that the NLS estimator converges in probability to a local minimum $\boldsymbol{\beta}^*$. Note that $f(\boldsymbol{x};\boldsymbol{\beta}^*)$ is, at most, an approximation to the conditional mean function.

7.3.2 Asymptotic Normality

Given that the NLS estimator $\hat{\boldsymbol{\beta}}_T$ is weakly consistent for some $\boldsymbol{\beta}^*$, we will sketch a proof that, with more regularity conditions, the suitably normalized NLS estimator is asymptotically distributed as a normal random vector.

First note that by the mean-value expansion of $\nabla_{\boldsymbol{\beta}} Q_T(\hat{\boldsymbol{\beta}}_T)$ about $\boldsymbol{\beta}^*$,

$$\nabla_{\boldsymbol{\beta}} Q_T(\hat{\boldsymbol{\beta}}_T) = \nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^*) + \nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}_T^\dagger)(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*),$$

where $\boldsymbol{\beta}_T^\dagger$ is a mean value between $\hat{\boldsymbol{\beta}}_T$ and $\boldsymbol{\beta}^*$. Clearly, the left-hand side is zero because $\hat{\boldsymbol{\beta}}_T$ is the NLS estimator and hence solves the first order condition. By [ID-2], the Hessian matrix $\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}_T^\dagger)$ is invertible, so that

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) = -[\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}_T^\dagger)]^{-1}\sqrt{T}\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^*).$$

The asymptotic distribution of $\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*)$ is therefore the same as that of the right-hand side.

Let $\boldsymbol{H}_T(\boldsymbol{\beta}) = \mathbb{E}[\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta})]$ and vec denote the operator such that $\mathrm{vec}(\boldsymbol{A})$ is the vector resulted from stacking all the column vectors of $\boldsymbol{A}$. By the triangle inequality,

$$\begin{aligned}&\left\|\mathrm{vec}\left[\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}_T^\dagger)\right] - \mathrm{vec}\left[\boldsymbol{H}_T(\boldsymbol{\beta}^*)\right]\right\| \\ &\quad \le \left\|\mathrm{vec}\left[\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}_T^\dagger)\right] - \mathrm{vec}\left[\boldsymbol{H}_T(\boldsymbol{\beta}_T^\dagger)\right]\right\| + \left\|\mathrm{vec}\left[\boldsymbol{H}_T(\boldsymbol{\beta}_T^\dagger)\right] - \mathrm{vec}\left[\boldsymbol{H}_T(\boldsymbol{\beta}^*)\right]\right\|.\end{aligned}$$

The first term on the right-hand side converges to zero in probability, provided that $\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta})$ also obeys a WULLN. As $\boldsymbol{\beta}_T^\dagger$ is a mean value between $\hat{\boldsymbol{\beta}}_T$ and $\boldsymbol{\beta}^*$, weak consistency of $\hat{\boldsymbol{\beta}}_T$ implies $\boldsymbol{\beta}_T^\dagger$ also converges in probability to $\boldsymbol{\beta}^*$. Under [ID-2], Q_T is twice continuously differentiable so that $\boldsymbol{H}_T(\boldsymbol{\beta})$ is continuous in $\boldsymbol{\beta}$. Thus, $\boldsymbol{H}_T(\boldsymbol{\beta}_T^\dagger) - \boldsymbol{H}_T(\boldsymbol{\beta}^*)$ also converges to zero in probability. Consequently, $\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}_T^\dagger)$ is essentially close to $\boldsymbol{H}_T(\boldsymbol{\beta}^*)$, in the sense that they differ by an $o_{\mathbb{P}}(1)$ term.

The result above shows that the normalized NLS estimator, $\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*)$, and

$$-\boldsymbol{H}_T(\boldsymbol{\beta}^*)^{-1}\sqrt{T}\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^*)$$

are asymptotically equivalent and hence must have the same limiting distribution. Observe that $\sqrt{T}\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^*)$ is a partial sum:

$$\sqrt{T}\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^*) = -\frac{2}{\sqrt{T}}\sum_{t=1}^{T} \nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)[y_t - f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)],$$

and hence obeys a CLT under suitable regularity conditions. That is,

$$(\boldsymbol{V}_T^*)^{-1/2}\sqrt{T}\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^*) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_k),$$

where

$$\boldsymbol{V}_T^* = \operatorname{var}\left(\frac{2}{\sqrt{T}}\sum_{t=1}^{T} \nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)[y_t - f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)]\right).$$

Then for $\boldsymbol{D}_T^* = \boldsymbol{H}_T(\boldsymbol{\beta}^*)^{-1}\boldsymbol{V}_T^*\boldsymbol{H}_T(\boldsymbol{\beta}^*)^{-1}$, we immediately obtain the following asymptotic normality result:

$$(\boldsymbol{D}_T^*)^{-1/2}\boldsymbol{H}_T(\boldsymbol{\beta}^*)^{-1}\sqrt{T}\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^*) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_k),$$

which in turn implies

$$(\boldsymbol{D}_T^*)^{-1/2}\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_k),$$

As in linear regression, asymptotic normality of the normalized NLS estimator remains valid when $\boldsymbol{D}_T^*$ is replaced by its consistent estimator $\hat{\boldsymbol{D}}_T$:

$$\hat{\boldsymbol{D}}_T^{-1/2}\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_k),$$

Thus, finding a consistent estimator for $\boldsymbol{D}_T^*$ is important in practice.

Consistent estimation of $\boldsymbol{D}_T^*$ is completely analogous to that for linear regression; see the discussion of Section 6.3. First observe that $\boldsymbol{H}_T(\boldsymbol{\beta}^*)$ is

$$\boldsymbol{H}_T(\boldsymbol{\beta}^*) = \frac{2}{T}\sum_{t=1}^{T} \mathbb{E}\big(\big[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)\big]\big[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)\big]'\big) - \frac{2}{T}\sum_{t=1}^{T} \mathbb{E}\big(\nabla^2_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)\big[y_t - f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)\big]\big),$$

which can be consistently estimated by its sample counterpart:

$$\hat{\boldsymbol{H}}_T = \frac{2}{T}\sum_{t=1}^{T}\big[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\big]\big[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\big]' - \frac{2}{T}\sum_{t=1}^{T}\nabla^2_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\hat{e}_t.$$

When $\epsilon_t = y_t - f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)$ are uncorrelated with $\nabla^2_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)$, $\boldsymbol{H}_T(\boldsymbol{\beta}^*)$ depends only on the expectation of the outer product of $\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)$ so that $\hat{\boldsymbol{H}}_T$ simplifies to

$$\hat{\boldsymbol{H}}_T = \frac{2}{T}\sum_{t=1}^{T}\big[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\big]\big[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\big]'.$$

This estimator is analogous to $\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'/T$ for $\boldsymbol{M}_{xx}$ in linear regression.

If $\boldsymbol{\beta}^* = \boldsymbol{\beta}_o$ so that $f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)$ is the conditional mean of y_t, we write $\boldsymbol{V}_T^*$ as

$$\boldsymbol{V}_T^o = \frac{4}{T}\sum_{t=1}^{T}\mathbb{E}\left(\epsilon_t^2\big[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}^o)\big]\big[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}^o)\big]'\right).$$

When there is conditional homoskedasticity: $\mathbb{E}(\epsilon_t^2|\mathcal{Y}^{t-1},\mathcal{W}^t) = \sigma_o^2$, $\boldsymbol{V}_T^o$ simplifies to

$$\boldsymbol{V}_T^o = \frac{4\sigma_o^2}{T}\sum_{t=1}^{T}\mathbb{E}\left(\big[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}_o)\big]\big[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}_o)\big]'\right),$$

which can be consistently estimated by

$$\hat{\boldsymbol{V}}_T = \frac{4\hat{\sigma}_T^2}{T}\sum_{t=1}^{T}\big[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\big]\big[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\big]',$$

with $\hat{\sigma}_T^2$ a consistent estimator for σ_o^2. Typically, $\hat{\sigma}_T^2 = \sum_{t=1}^{T}\hat{e}_t^2/T$. In the NLS context, $\sum_{t=1}^{T}\hat{e}_t$ is non-zero in general so that we may also compute $\hat{\sigma}_T^2$ as

$$\hat{\sigma}_T^2 = \frac{1}{T}\sum_{t=1}^{T}(\hat{e}_t - \bar{\hat{e}})^2,$$

where $\bar{\hat{e}} = \sum_{t=1}^T \hat{e}_t/T$. Combining $\hat{\boldsymbol{V}}_T$ and $\hat{\boldsymbol{H}}_T$ we have

$$\hat{\boldsymbol{D}}_T = \hat{\sigma}_T^2 \left(\frac{1}{T} \sum_{t=1}^T \left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t; \hat{\boldsymbol{\beta}}_T)\right] \left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t; \hat{\boldsymbol{\beta}}_T)\right]' \right)^{-1}.$$

This estimator is analogous to the OLS covariance matrix estimator $\hat{\sigma}_T^2(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}$ for linear regressions.

When there is conditional heteroskedasticity such that $\mathbb{E}(\epsilon_t^2|\mathcal{Y}^{t-1}, \mathcal{W}^t)$ are functions of the elements of $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$, $\boldsymbol{V}_T^o$ can be consistently estimated by

$$\hat{\boldsymbol{V}}_T = \frac{4}{T} \sum_{t=1}^T \hat{e}_t^2 \left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t; \hat{\boldsymbol{\beta}}_T)\right] \left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t; \hat{\boldsymbol{\beta}}_T)\right]',$$

so that

$$\hat{\boldsymbol{D}}_T = \left(\frac{1}{T} \sum_{t=1}^T \left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t; \hat{\boldsymbol{\beta}}_T)\right] \left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t; \hat{\boldsymbol{\beta}}_T)\right]' \right)^{-1} \hat{\boldsymbol{V}}_T \left(\frac{1}{T} \sum_{t=1}^T \left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t; \hat{\boldsymbol{\beta}}_T)\right] \left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t; \hat{\boldsymbol{\beta}}_T)\right]' \right)^{-1}.$$

This is White's heteroskedasticity-consistent covariance matrix estimator for nonlinear regressions.

As discussed earlier, the probability limit $\boldsymbol{\beta}^*$ of the NLS estimator is typically a local minimum of $\mathbb{E}[Q_T(\boldsymbol{\beta})]$ and hence not $\boldsymbol{\beta}_o$ in general. In this case, $\{\epsilon_t\}$ is not a martingale difference sequence with respect to $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$, and $\boldsymbol{V}_T^*$ must be estimated using a Newey-West type estimator; see Exercise 7.7.

7.4 Hypothesis Testing

For testing linear restrictions of parameters, we again consider the null hypothesis

$$\mathrm{H}_0 \colon \boldsymbol{R}\boldsymbol{\beta}^* = \boldsymbol{r},$$

where $\boldsymbol{R}$ is a $q \times k$ matrix and $\boldsymbol{r}$ is a $q \times 1$ vector of pre-specified constants.

The Wald test now evaluates the difference between the NLS estimates and the hypothetical values. When the normalized NLS estimator, $T^{1/2}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o)$, has an asymptotic normal distribution with the asymptotic covariance matrix $\boldsymbol{D}_T^*$, we have under the null hypothesis that

$$(\Gamma_T^*)^{-1/2}\sqrt{T}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}) = (\Gamma_T^*)^{-1/2}\sqrt{T}\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) \xrightarrow{D} N(0, \boldsymbol{I}_q).$$

where $\Gamma_T^* = \boldsymbol{R}\boldsymbol{D}_T^*\boldsymbol{R}'$. Let $\hat{\boldsymbol{D}}_T$ be a consistent estimator for $\boldsymbol{D}_T$. Then, $\hat{\Gamma}_T = \boldsymbol{R}\hat{\boldsymbol{D}}_T\boldsymbol{R}'$ is also consistent for Γ_T^*. It follows that the Wald statistic is

$$\mathcal{W}_T = T(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})\hat{\Gamma}_T^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})' \xrightarrow{D} \chi^2(q),$$

which is of the same form as the Wald statistic in linear regressions.

More generally, one may test for the nonlinear restriction of the form

$$\mathrm{H}_0 \colon \boldsymbol{r}(\boldsymbol{\beta}^*) = \boldsymbol{0},$$

where $\boldsymbol{r}$ now is a $\mathbb{R}^q$-valued nonlinear function that is continuously differentiable. It is natural to consider basing the Wald test on $\boldsymbol{r}(\hat{\boldsymbol{\beta}}_T)$. First note that linearization of $\boldsymbol{r}(\hat{\boldsymbol{\beta}}_T)$ about $\boldsymbol{\beta}^*$ yields

$$\boldsymbol{r}(\hat{\boldsymbol{\beta}}_T) = \boldsymbol{r}(\boldsymbol{\beta}^*) + [\nabla_{\boldsymbol{\beta}}\boldsymbol{r}(\boldsymbol{\beta}^*)]'(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) + o_{\mathbb{P}}(1).$$

Letting $\boldsymbol{R}(\boldsymbol{\beta}^*) = [\nabla_{\boldsymbol{\beta}}\boldsymbol{r}(\boldsymbol{\beta}^*)]'$, a $q \times k$ matrix, we have under the null hypothesis that

$$\boldsymbol{r}(\hat{\boldsymbol{\beta}}_T) = \boldsymbol{R}(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) + o_{\mathbb{P}}(1).$$

Similar as before,

$$(\Gamma_T^*)^{-1/2}\boldsymbol{r}(\hat{\boldsymbol{\beta}}_T) = (\Gamma_T^*)^{-1/2}\sqrt{T}\boldsymbol{R}(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) \xrightarrow{D} N(0, \boldsymbol{I}_q).$$

where in this case $\Gamma_T^* = \boldsymbol{R}(\boldsymbol{\beta}^*)\boldsymbol{D}_T^*\boldsymbol{R}(\boldsymbol{\beta}^*)'$. This immediately suggests the following Wald statistic:

$$\mathcal{W}_T = T\boldsymbol{r}(\hat{\boldsymbol{\beta}}_T)'\hat{\Gamma}_T^{-1}\boldsymbol{r}(\hat{\boldsymbol{\beta}}_T) \xrightarrow{D} \chi^2(q),$$

where $\hat{\Gamma}_T = \boldsymbol{R}(\hat{\boldsymbol{\beta}}_T)\hat{\boldsymbol{D}}_T\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T)'$ is consistent for Γ_T^*.

A well known drawback of the Wald test is that it is not invariant with respect to different forms of a nonlinear hypothesis. For example, consider two equivalent hypotheses: $\beta_1\beta_2 = 1$ and $\beta_1 = 1/\beta_2$. The Wald test for the former is based on $\hat{\beta}_1\hat{\beta}_2 - 1$, whereas the Wald test for the latter is based on $\hat{\beta}_1 - \hat{\beta}_2^{-1}$. It turns out that these two tests perform very differently; see e.g., Gregory and Veall (1985) and Phillips and Park (1988). In particular, the Wald test for $\beta_1 = 1/\beta_2$ rejects far too often when the null hypothesis is indeed correct (i.e., the empirical size is much larger than the nominal size). Moreover, these two tests result in conflicting conclusions quite often. Hence, the inferences from testing nonlinear hypotheses should be drawn with care.

Exercises

7.1 Suppose that $Q_T(\boldsymbol{\beta})$ is quadratic in $\boldsymbol{\beta}$:

$$Q_T(\boldsymbol{\beta}) = a + \boldsymbol{b}'\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{C}\boldsymbol{\beta},$$

where a is a scalar, $\boldsymbol{b}$ a vector and $\boldsymbol{C}$ a symmetric, positive definite matrix. Find the first order condition of minimizing $Q_T(\boldsymbol{\beta})$ and the resulting solution. Is the OLS criterion function (3.2) quadratic in $\boldsymbol{\beta}$?

7.2 Let $\hat{\epsilon}_t = y_t - \hat{y}_t$ denote the t^{th} NLS residuals. Is $\sum_{t=1}^T \hat{\epsilon}_t$ zero in general? Why or why not?

7.3 Given the nonlinear specification of the CES production function

$$\ln y = \ln \alpha - \frac{\lambda}{\gamma} \ln\left[\delta L^{-\gamma} + (1-\delta)K^{-\gamma}\right] + e,$$

find the second order Taylor expansion of $\ln y$ around $\gamma = 0$. How would you estimate this linearized function and how can you calculate the original parameters α, γ, δ and λ?

7.4 Suppose that $\mathbb{E}[Q_T(\boldsymbol{\beta})]$ is a continuous function on the compact parameter space Θ_1 such that $\boldsymbol{\beta}_o$ is its unique, global minimum. Also suppose that the NLS estimator $\hat{\boldsymbol{\beta}}_T$ is such that

$$\mathbb{E}[Q_T(\hat{\boldsymbol{\beta}}_T)] = \inf_{\Theta_1} \mathbb{E}[Q_T(\boldsymbol{\beta})].$$

Prove that when Q_T has a WULLN effect, then $\hat{\boldsymbol{\beta}}_T$ converges in probability to $\boldsymbol{\beta}_o$.

7.5 Apply Theorem 7.2 to discuss the consistency property of the OLS estimator for the linear specification $y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t$.

7.6 Let $\epsilon_t = y_t - f(\boldsymbol{x}_t;\boldsymbol{\beta}_o)$. If $\{\epsilon_t\}$ is a martingale difference sequence with respect to $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$ such that $\mathbb{E}(\epsilon_t^2 \mid \mathcal{Y}^{t-1}, \mathcal{W}^t) = \sigma_o^2$, state the conditions under which $\hat{\sigma}_T^2 = \sum_{t=1}^T \hat{e}_t^2/T$ is consistent for σ_o^2.

7.7 Let $\epsilon_t = y_t - f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)$, where $\boldsymbol{\beta}^*$ may not be the same as $\boldsymbol{\beta}_o$. If $\{\epsilon_t\}$ is not a martingale difference sequence with respect to $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$, give consistent estimators for $\boldsymbol{V}_T^*$ and $\boldsymbol{D}_T^*$.

References

Amemiya, Takeshi (1985). *Advanced Econometrics*, Cambridge, MA: Harvard University Press.

Davidson, Russell and James G. MacKinnon (1993). *Estimation and Inference in Econometrics*, New York, NY: Oxford University Press.

Bierens, Herman J. (1994). *Topics in Advanced Econometrics*, New York, NY: Cambridge University Press.

Gallant, A. Ronald (1987). *Nonlinear Statistical Inference*, New York, NY: John Wiley & Sons.

Gallant, A. Ronald and Halbert White (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*, Oxford, UK: Basil Blackwell.

Gregory, Allan W. and Michael R. Veall (1985). Formulating Wald tests of nonlinear restrictions, *Econometrica*, **53**, 1465–1468.

Kuan, Chung-Ming and Halbert White (1994). Artificial neural networks: An econometric perspective, *Econometric Reviews*, **13**, 1–91.

Phillips, Peter C. B. and Joon Y. Park (1988). On the formulation of Wald tests of nonlinear restrictions, *Econometrica*, **56**, 1065–1083.